国防电子信息技术丛书

军事信息技术基础

Fundamental of Military Information Technology

高秀峰　齐剑锋　　主编

崔　静　李　芳　刘爱珍

王　路　王　帅　王寅龙　　参编

（排名不分先后）

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书从信息流程的角度，介绍信息的采集获取、交换传输、存储管理、加工处理，以及安全防护的相关基础技术，使读者初步了解信息技术的范畴和基本内容，为以后深入学习和有效应用信息技术打下良好的基础。

全书共分八章。第1章绪论，主要介绍信息、信息技术、军事信息技术相关的概念和基础知识；第2章信息获取技术，主要介绍光电、雷达、声波、地面传感器、卫星定位等信息获取技术的原理和特点；第3章信息传输与交换技术，介绍通信的基本概念，光纤、微波、卫星等信息传输技术，以及信息交换和信息网络技术；第4章信息存储与管理技术，主要介绍信息资源的分类编码方法、常用存储设备，以及数据库、数据仓库和数据容灾技术；第5章信息加工技术，主要介绍数据挖掘、模式识别、信息融合、数据可视化等高级数据处理技术的基本概念和基本方法；第6章信息服务技术，主要介绍情报信息和网络信息的分类组织、标引整序、查询检索，以及提供服务等技术；第7章信息安全技术，介绍信息安全、网络安全相关技术基础知识，主要包括加密技术、信息隐藏技术、网络安全技术等；第8章典型军事信息系统，介绍指挥自动化系统的构成及典型装备，以说明信息获取、传输、管理、加工等技术的应用情况。

本书可作为军事院校本科生信息技术基础课程的教材，也可供对信息技术感兴趣的读者参考。

图书在版编目（CIP）数据

军事信息技术基础/高秀峰，齐剑锋主编. —北京：电子工业出版社，2017.9
（国防电子信息技术丛书）
ISBN 978-7-121-30770-6

Ⅰ. ①军… Ⅱ. ①高… ②齐… Ⅲ. ①信息技术－应用－军事－高等学校－教材 Ⅳ. ①E919

中国版本图书馆CIP数据核字（2016）第322899号

策划编辑：马 岚
责任编辑：马 岚　　特约编辑：赵晓温
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787×1092　1/16　印张：17　字数：490千字
版　　次：2017年9月第1版
印　　次：2017年9月第1次印刷
定　　价：59.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：classic-series-info@phei.com.cn。

前　言

随着世界新军事变革的深入发展和信息时代的到来，军队战斗力的要素、标准和内涵都发生了深刻变化，信息能力在战斗力生成中越来越具有主导作用，信息化武器装备成为战斗力的关键物质因素，基于信息系统的体系作战能力已成为战斗力的基本形态。然而，武器装备归根结底是由人来使用的，信息化武器装备对使用人员和指挥员都提出了更高的信息素质要求。

为了跟上信息化武器装备发展的形势，我军提出提高军队学员综合素质的要求，其中信息素质是综合素质的重要组成部分。很多院校都在探索如何提高军校学员的信息素质，有的院校设置了信息技术基础一类的课程，目的是讲授信息技术的基础知识，培养学员信息意识，提高学员信息分析和运用能力，加深学员对信息的价值及其中存在的风险的认识，从而为学员今后更好地操作运用信息化装备、指挥保障信息化作战打下良好的基础。

然而，信息技术不是一门技术，而是一个技术群，覆盖范围很广，内容极多，而学员的学习时间是有限的，特别是非信息技术专业的学员，其时间更为有限。如何在有限的时间内，使学员掌握所需的信息技术知识，训练一定的信息技术技能，培养敏锐的信息意识，是一个困难而又重要的课题。这就要求对信息技术各方面的知识精心地挑选和组织，准确地把握合适的深度，以适于学员学习和掌握。

目前信息技术基础教材很多，但大多以Office等常用软件操作和简单编程知识作为主要内容，和军校学员所需信息技术有较大差别，因此我们编写了本教材，这也是对提高学员信息素质教学的一次探索。本教材编写时注意把握两个方面：（1）不要只泛泛地介绍概念；（2）不要介绍太深的技术内容。作为基础，应有利于没有太多基础的学生入门学习。本教材内容着重于在军事上有重要应用的信息技术，即军事信息技术，根据信息技术的作战运用规律来取舍，按照军事信息的采集获取、交换传输、加工处理、存储管理、应用服务的流程来编排，同时考虑贯穿流程的安全保密工作，用精心选择的知识点描绘出一个完整的信息技术军事运用轮廓，为学员提高信息化装备的灵活运用能力和今后进一步深入学习打下良好的基础。

全书共分八章。第1章和第3章主要由刘爱珍编写（其中3.2.3节由王帅编写），第2章主要由王路编写（其中2.6节由王帅编写），第4章由崔静编写，第5章由齐剑锋编写，第6章由王寅龙编写，第7章由高秀峰编写，第8章由李芳编写。高秀峰、齐剑锋统校了全稿。

由于时间紧迫，笔者水平有限，书中难免有不当之处，欢迎读者批评指正。

《军事信息技术基础》教材编写组

二〇一六年十二月于石家庄

目　　录

第1章　绪　　论

当前，人类正逐步迈向信息社会，信息的开发利用水平空前提高，各类信息技术得到快速发展并广泛应用，对科技发展、经济增长、社会进步和战争胜利的作用日益增强。近年来，以信息化为首要特征的世界新军事变革，正在把机械化军事形态改造成信息化军事形态。加快军队信息化建设，推动军队向信息化转型已成为世界各国的普遍选择。随着我军信息化建设步伐的加快，信息化范围不断拓展，逐步向纵深发展，已进入全面建设阶段。学习信息基本知识，掌握信息技术，提高信息素质，已成为信息时代每个军人的基本功课。本章就从信息的基本知识谈起，以使我们对信息及信息技术有个概要的了解。

1.1　信息

1.1.1　人类与信息

人类的生产生活一时一刻都离不开信息，人类对信息的认识经历了一个不断深化的长期过程。纵观历史，人类对信息的感知、传递、处理和利用的能力经历了五次跃升，其主要标志分别是语言的产生、文字的诞生、印刷术的发明、电磁波的利用和计算机的出现（见图1.1）。

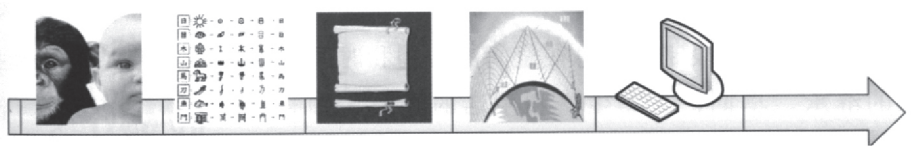


图1.1　人类对信息的感知、传递、处理和利用的能力的五次跃升

1. 第一次跃升——语言的产生

亿万年来，无论是否被感知、被发现，信息一直伴随着事物的运动存在于宇宙中。当地球上出现生命后，信息开始被动物的眼、鼻、口、耳等感觉器官所感知、所察觉，并最终通过动物的神经传递到大脑，形成反映而被利用。在人类进化过程中，随着劳动的复杂性不断提升，相对简单的表情、鸣叫和动作已不足以描述复杂的环境和表达丰富的情感。人类发出的音调出现了高、低、粗、细的变化，由简单到复杂，由零星、断续到逻辑连贯，通过不断地磨练和积累，促使了发声器官的进化和完善，人类终于创造出了语言，实现了人类感知、传递、处理和利用信息的能力的第一次跃升。

语言的产生标志着人类信息活动的范围和效率有了质的跃升，并大大促进了人类大脑的发展，增强了人的表达能力、理解能力、抽象能力和推理能力，最终使人与动物彻底区分开来，

拉开了人类文明的序幕。因此，“语言”成为人类顺应自然、利用自然、改造自然的第一个信息平台。

2. 第二次跃升——文字的诞生

在人类信息活动当中，语音是最早的信息载体。早期，人类生产和生活的经验、知识，唯有通过氏族部落长者向晚辈言传身教的方式，代代相传，承袭下去。随后出现了结绳记事（见图1.2），人类通过绳结的大小、样式、颜色等来表达自己的意愿，记录人类的历史。由于生产活动的进步和物质财富的积累，以及贵族权杖的出现和宗教礼仪活动的日益频繁，人们便产生了要把更多、更复杂的事物记录下来的要求。于是，出现了最早的刻划符号，这标志着文字形态开始萌芽。在距今五六千年以前的我国黄河流域的仰韶文化、大汶口文化等新石器人类遗址中，已经出现了刻画在陶器上的象形文字。从语言发展到文字，实现了人类感知、传递、处理和利用信息的能力的第二次跃升。

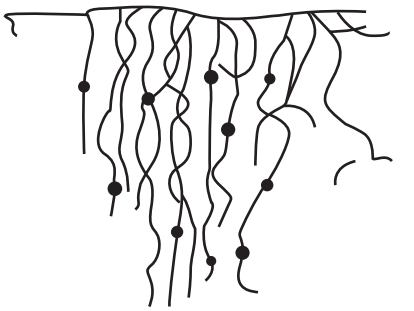


图1.2 结绳记事

文字的产生揭开了人类历史的新篇章，这是一次信息载体和传播手段的重要突破，从此人类可以将其生产生活方式、对自然的认识体会等记载下来，真正有了“确凿”的历史，突破了原来的时空限制，从而使人类获取的知识得到传承和积累，为人类智慧水平的不断提高提供了条件。

3. 第三次跃升——印刷术的发明

随着文字的产生、发展和演变，人类不断创造和发明新的记载材料和记载方法。文字记载材料经历了从石器、泥板、甲骨到铜器、简牍、绢帛的演变，后来出现了划时代的材料——纸张。纸张具有材料便宜、制作成本低廉和携带方便省力、便于保存的优点，因而很快成为一种大量保存和传递信息的载体。

同时，文字记录载体的变革也推动着信息记录方法的创新，印刷术应运而生。最先发明的是刻版印刷，随后出现了活字印刷术。印刷术的不断进步，产生了可容纳更多信息量的报纸、书籍、杂志，进而极大地提高了人类的信息交流水平。可以说，造纸术、印刷术的发明，使知识的积累和传播突破了历史、时空和地域界限，人类信息传递的速度和范围急剧扩展，人类信息的存储能力显著加强，并初步具备了广泛传播信息和共享信息的条件，从而实现了人类感知、传递、处理和利用信息的能力的第三次跃升。

4. 第四次跃升——电磁波的利用

在不断改进信息记录方式的同时，人类从来都没有放弃对信息时效性的追求。西周时期就有利用烽火传递信息的记载，后来又出现了驿站传递信息，但这些还远远不能很好地解决信息传递的时效性问题。到18世纪末期，法国人夏普发明了横木通信机，远距离信息传递的时效性得到了较大提高，但是这种通信方式不仅用人多，而且费用高昂。

19世纪初，人们发现电磁波可以承载信息。1837年，美国人莫尔斯发明并建成了电报线路。1876年，美国科学家贝尔发明了电话，由此带来了电报、电话等有线通信手段的发展。1887年，德国科学家赫兹利用火花隙激励一个环状天线，用另一个带缝隙的天线进行接收，证实了电磁

波真实地存在于空气中，由此架起了电磁波从有线通向无线的桥梁。人类利用电磁波传递信息拉近了世界的距离，使人们传递信息的能力迅速提高，同时也推动了科学技术迅猛发展，这便是人类历史上感知、传递、处理和利用信息的能力的第四次跃升。

这次跃升使传播信息的手段和载体、方式和方法都发生了质的飞跃。它不仅使语言、文字信息编码化，而且极大地提高了时空利用率。

5. 第五次跃升——计算机的出现

电磁波的应用又一次把信息传播的手段向前推进，但仍存在着很大局限性。20世纪中叶，科学技术迅猛发展，知识呈指数级增长，新事物、新概念层出不穷，使人目不暇接。人们普遍感到，不但需要更便捷和高效的信息传播手段，而且更需要能够辅助人类大脑进行信息分析和处理的工具。于是，电子计算机就应运而生了，从而带来了人类感知、传递、处理和利用信息的能力的第五次跃升。

这是一次同时包括了信息传播手段与信息处理手段的全面跃升，是综合了光、电、磁、声等多种类的多频段信息载体和传递工具的变革，是人类发展迈出的重要一步。电子计算机的出现对人类社会发展的影响是全方位的，它极大地加速了信息处理和交互，大大缓解和消除了人类传播信息在时空上的限制，世界由此成为“地球村”。

在科学技术迅猛发展的推动下，人类感知、传递、处理、利用信息的能力将不断出现新的跃升。

1.1.2　信息的含义

从某种意义上说，人类信息活动的演进和人类信息能力的发展伴随着整个人类的进化。那么什么是信息呢?

不同学者对信息有不同的解释，较具代表性的有以下几种：

(1) 1928年，哈特莱：信息是指有新内容、新知识的消息。

(2) 1948年，香农：信息是用以消除不确定性的东西。

(3) 1948年，维纳：信息是人们在适应外部世界、控制外部世界的过程中，与外部世界交换的内容。

(4) 1975年，朗高：信息是反映事物的形成、关系和差别的东西，它包含在事物的差异中，而不是事物本身。

这些解释的不同，实际上是由于看待信息的角度不同。哈特莱从价值的角度来解释，香农从通信的角度来解释，维纳从控制的角度来解释，朗高从认知的角度来解释。

一般认为，香农给出的信息定义更接近信息的本质，应用得更多一些。所以，本书中采用香农给出的信息定义。信息的度量方法和信息三要素的说法都来自香农的信息论。

但是，为便于理解和使用，本书综合几种资料的说法，给出一个更通俗的解释：信息是指客观事物存在的方式或运动的状态，它通过一定载体反映出来，体现客观事物相互联系的程度及规律。

所谓“客观事物”，泛指一切可能的研究对象，包括客观世界的一切事物和现象，既包括有形事物，如：风、雨、雷、电，高山、楼宇等；又包括无形事物，如：人们的思想、认识、情绪、策划、方法等。所谓“运动”，泛指一切意义上的变化，包括机械运动、物理运动、化学运动、生物运动、思维运动和社会运动等。所谓“运动的状态”，是指事物运动在空间上所展示的形状、态势及其在时间上的变化。

信息的分类多种多样。按来源，信息可分为自然信息和社会信息；按逻辑，信息可分为真实信息、虚假信息、不定信息等；按作用，信息可分为有用信息、无用信息、干扰信息等；按载体，信息可分为电子信息、光学信息、生物信息等；按应用领域，信息可分为政治信息、经济信息、军事信息、科技信息、文化信息等。

值得注意的是，在日常用语中，信息经常与消息、信号、数据、情报和知识等比较相近的概念交替使用，有时人们甚至把它们当成一回事，但实际上它们是有区别的，如下所示。

- 消息是由具体文字、符号或语音所表达的已发生的某个事件。消息是信息的外壳，信息是消息的内核。一个消息的产生，可能带来信息，也可能不带来信息。也就是说，不同消息中所含的信息量是不同的。
- 信号是用来承载信息的物理载体，信息是事物运动的状态和方式。
- 数据是信息的一种记录形式，但不是唯一的记录形式，除此之外，信息还可以通过文字、图形、语言等各种形式记录。
- 情报是一类特殊的信息，是信息集合的一个子集，任何情报都是信息，但并非所有信息都是情报。
- 知识是关联起来的信息，是信息加工的产物，是一种高级形式的信息。比如，“天冷了”是信息，“大雁往南飞了”是信息，“天冷了，大雁就要往南飞了”就是知识，它是经过人的大脑加工的结果。任何知识都是信息，但并非任何信息都是知识。

从上述相近的概念区分中可以看出，信息有低级和高级之分。低级信息是人们无须花费多大力气就能收集到的事实性知识，只是部分事物的比较片面的反映，例如广告、电影、电视节目预告等；高级信息是人们经过一番努力，进行深入加工处理而收集到的知识，例如公司的年终报表等；更高级、更有价值的信息是那些与创造发明有关的包含智慧结晶的信息，例如机器的发明、国家的战略规划、重大工程项目的决策等。

信息与物质、能量一起构成了人类社会赖以生存和发展的三大基石，是一种重要的战略资源。物质为人类提供材料，能量为人类提供动力，信息则为人类提供知识和智慧。信息是认识世界和改造世界的首要条件，没有信息，人类就不可能认识世界，更不可能改造世界。因此，明确信息的度量方式，熟悉信息的要素、基本形态、主要特征和基本作用等方面，对于我们更有效地掌握、认知信息，更有效地认识世界和改造世界，有着重要的意义。

1.1.3 信息三要素

信源、信宿和信道是信息三要素。

信源，又称信息源，是信息的发源地，或者说是信息的出处。信源大体分为三大类：(1) 来自自然界，包括天体、地理、生物等方面的信息；(2) 来自社会，包括人类社会的生产、经济、军事等方面的动态与情报；(3) 他人的知识，包括古今中外流传下来的知识及专家学者的经验。

信宿，是信息的归宿，是接受信息者对信息判断后做出的处理结果。信宿决定信息的价值。信息被有关者获取后，通过加工处理、正确理解和正确使用，才能真正发挥作用。因此，信息获取者要对信息进行筛选分类，综合分析，分清哪些是有用信息，哪些是无用信息甚至假信息，以便利用有价值的信息，摒弃无价值的信息。

信道，传递信息的通道，是信源与信宿之间联系的纽带。信道有自然信道、人体的本能信道和技术信道。空气、风、水等是自然信道；人体的四肢、五官等感觉器官是本能信道；无线电通信、计算机网络等是技术信道。

1.1.4　信息的度量

信息与消息有着不可分割的内在联系，不同消息中所含的信息量是不同的，那么应该如何衡量信息量呢？香农给出了方法：消息中含信息量的大小是由它消除的不确定程度决定的。

消息中含有的信息量与消息发生的概率紧密相关。某消息出现的概率越小，则其包含的信息量越大；某消息出现的概率越大，则其包含的信息量越小。必然事件的发生不带来任何信息。独立事件的发生可看成消息的发生，若干事件的联合发生也可看成消息的发生，该消息带来的信息应与其中各个事件有关。如果消息由符号组成，而各符号又被看成独立发生的，则多符号联合的消息的发生概率呈指数规律减小。

综合以上情况可知，消息中所含的信息量与消息发生的概率有以下关系：

(1) 发生的概率越小，消息中所含的信息量越大，消息带来的信息量与消息发生的概率成反比；

(2) 联合消息发生的概率呈指数规律减小，或呈指数规律增加。

受这些规律支配，对信息量定义如下：

设消息x发生的概率为$P(x)$，则该消息带来的信息量定义为

$$I(x)=\log_a(1/P(x))=-\log_a P(x) \tag{1.1}$$

其中，取对数可使原指数规律变得平稳，便于表达。若对数的底取2，则I的单位为bit；若对数的底取e，则I的单位为Net；若对数的底取10，则I的单位为Hatle。bit是最常用的单位。

例1.1　计算等概率发生的离散消息的信息量。

解：设信源在每个时刻发生的消息非0即1，此即二进制符号消息，它们出现的概率都是1/2，则其所带的信息量为

$$I_2=\log_2 2=1 \quad (\text{bit})$$

即二进制符号的每个码元带来1 bit的信息量。

若信源发出的消息为M进制，M进制符号消息出现的概率为

$$P(0)=P(1)=\cdots=P(M)=1/M$$

则其所带的信息量为

$$I_M=\log_2 M \quad (\text{bit})$$

且若$M=2^k$（$K=1, 2, 3, \cdots$），则

$$I_M=\log_2 2^k=K \quad (\text{bit})$$

很显然，一个消息的用bit表示的信息量和表示或存储这个消息所需的二进制位数是相等的。我们知道了消息的发生概率，就可以计算其用bit表示的信息量，这就是要表示或存储该消息所需占用的bit数。

式（1.1）所表示的是一个符号所带的信息量，如果想知道多个符号所带的信息量，只需把所有符号的信息量求和即可。根据定义，信息量代表了符号或符号串的不确定性。

很多情况下，我们希望了解信源的不确定性特征，也就是信源发出的每个符号平均消除的不确定性，也就是每个符号平均的信息量。在信息论中，这就是信源的“熵”，用H表示，如果信源用x表示，则该信源的熵用$H(x)$表示。

对于等概率信源，每个符号的信息量是相等的，所以每个符号平均的信息量就等于其中一个符号的信息量。对于例1.1中发出两个符号的信源，其熵为

$$H(x) = I_2 = \log_2 2 = 1 \quad \text{(bit)}$$

对于发出M进制符号的信源x，其熵为

$$H(x) = I_M = \log_2 M \quad \text{(bit)}$$

且若$M = 2^k$（$K = 1, 2, 3, \cdots$），则

$$H(x) = I_M = \log_2 2^k = K \quad \text{(bit)}$$

上面说的是等概率信源的熵的计算方法。在非等概率情形下，则需要把熵的计算方法进行一般化推广。

设信源x的各个符号为：x_i(i = 1, 2, 3, …, N)，其概率分别为$P(x_i)$，则定义熵为该信息源各符号的统计平均信息量，以$H(x)$表示。

$$H(x) = \sum_{i=1}^{N} P(x_i) I(x_i) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i) \tag{1.2}$$

例1.2 信源发出4个符号0、1、2和3，信源发出它们的概率分别为3/8、1/4、1/4和1/8，求该信源的熵。

解：

$$\begin{aligned} H(x) &= \sum_{i=1}^{4} P(x_i) I(x_i) \\ &= \frac{3}{8} \log_2 \left(\frac{3}{8}\right) + \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 + \frac{1}{8} \log_2 8 \\ &= 1.906 \text{ bit} \end{aligned}$$

信源的熵用bit为单位表示时，它就等于表示或存储该信源的每个符号所需的二进制位数的平均值。

1.1.5 信息的主要特征

总体而言，信息主要具有可识别性、共享性、可伪性、时效性，以及价值相对性等特征。

1. 信息的可识别性

信息可以通过某种媒介，以某种方式被人类所感知，人类进而可掌握信息所反映的客观事物的状态和运动方式，这就是信息的可识别性。目前，人类能够接受和使用的信息，只是无限丰富信息中的一部分，还有许多信息尚未被人们所认识，但这并不是说这些信息不可识别，只是受科学技术水平所限，人类尚未了解承载该信息的媒介和方式。

随着科学技术的发展，人类感知信息的手段和能力将不断提升，获取的信息也将越来越多。例如，人类凭借肉眼可通过明暗或颜色区分不同的物体，这靠的是物体反射或者辐射的可

见光，地球上的物体大多数是不发光的，所以到了晚上，人类就看不清东西了，于是就会减少活动。后来人类发明了热成像仪，它能够获取物体辐射出的红外射线，并转换成可见光图像，所有的物体每时每刻都在辐射出红外线，人们借助热成像仪，就能在晚上看到各种物体，在必要的情况下，人们就可以增加晚上的活动。

2. 信息的共享性

信息可以被无限制地复制、传播或分配给众多用户，并能在这个过程中保持低损耗甚至无损耗，这就是信息的共享性。信息的共享性突出表现在两个方面：(1) 信息脱离所反映的事物而独立存在并附于其他载体，而载体在空间上的位移，使信息能够在不同空间和不同对象之间进行传递；(2) 信息不像水、石油、货币这些物质遵循守恒原则（即总量固定、与他人共享必然带来损耗甚至丧失），信息可以被大量复制、广泛传递。

例如，甲有一个苹果，乙也有一个苹果，那么甲乙互相交换之后，甲还只有一个苹果，乙也还只有一个苹果；但是如果甲有一条信息，乙也有一条信息，那么甲乙互相交换之后，甲就有了两条信息，乙也有了两条信息。信息的共享性特征对人类具有特别重要的意义。在军事行动中，其意义集中体现在共享战场信息是实施信息化联合作战的重要保障。

3. 信息的可伪性

信息能够被人类主观地加工、改造，进而产生畸变。同时，通过一定方式和手段，也可使人类对信息产生失真甚至错误的理解认识，这就是信息的可伪性。信息具备可伪性的原因在于信息不是事物本身，人们主观片面理解信息，或根据自己的意图，有意或无意地对信息的内容及负载信息的载体施加影响，就有可能使信息无法真实反映事物本身及其运动状态的原貌。

1944年，盟军在诺曼底登陆之前成功地进行了信息欺骗和信息封锁，造成德军对盟军登陆地点的判断失误。一系列信息欺骗行动有效地掩护了盟军的主攻方向，当盟军已经在诺曼底抢滩登陆时，深受欺骗的希特勒还担心盟军会在加莱地区实施更大规模的登陆作战。三国时期，诸葛亮用“空城计”吓退司马懿十五万大军也是同样的道理。

4. 信息的时效性

信息的价值会随时间的推移而改变，这就是信息的时效性。由于事物本身在不断发展变化中，因此信息必须随之变化才能准确反映事物的运动状态和状态的变化方式。信息被传递后就会脱离事物，原信息便不能反映事物的新的运动状态和状态变化方式，效用会逐渐降低，甚至完全丧失。

5. 信息的价值相对性

信息的价值相对性是指同样的信息对于不同的人具有不同的价值。这是由于信息的价值与信息接受者的观察能力、想象能力、思维能力、注意力和记忆力等智力因素密切相关，同时也依赖于他的知识结构和知识水平。

街口的信号灯变化对色盲患者是没有价值的无用信息，然而对正常人却至关重要，莎士比亚说“一千个观众眼里会有一千个哈姆雷特”，就是这个道理。

1.1.6　信息的基本作用

人类活动的全部目的是认识世界并改造世界。因此，人们获取信息，就是掌握和理解有关客观事物的运动状态和变化方式，把握其中的规律，积累并创新知识，消除各种各样的不确定

性，更加准确、高效地认识世界和改造世界。从这个角度来看，信息的基本作用主要包括反映作用、联系作用和启迪作用。

1. 信息的反映作用

信息的反映作用是指信息能够直接或间接地反映事物的存在方式或运动状态。其最简单和直接的形式表现为：信息能够对人类的视觉、听觉、嗅觉、触觉等感觉器官造成刺激。当我们看到苹果红了，就知道苹果熟了，并且可以食用了。苹果红了这个信息，就反映出了苹果已经生长成熟的状态。

2. 信息的联系作用

信息的联系作用是指任何系统，无论是生命系统还是非生命系统，其相互关联与交流都要以信息为中介，其间物质和能量的变化、运动和交换也都以信息的联系为先导。如果没有信息的联系作用，则任何一个系统的正常运转都不可能实现，系统将陷入混乱和无序。对任何个体而言，亦是如此，即使是人类的意识活动，包括人群之间的思想和情感交流，也必须通过信息联系才能实现。

信息的联系作用广泛存在于各个领域。存在于生命过程、感觉器官与外部世界的联系、神经中枢与各部分器官的联系、亲代把性状特征遗传给子代等生物领域，存在于通信系统、控制系统、火箭和导弹的制导系统、电子计算机系统等技术领域，也存在于生产过程、经济管理、文学艺术、历史考古等社会领域。

3. 信息的启迪作用

信息的启迪作用是指信息能够开导或启发人类进行更高效或更具目的性的活动。信息是人类认识世界的一扇大门，其中所蕴含的意义可以直接被人类了解，也可以通过人工装置或者生物间接地被人类了解，并产生启迪作用，进而影响或控制人类的思维和行为，更加深刻地改造世界。人类的生存依赖于大自然，人类的发展浸润在大自然所散发出的庞大信息中。正是由于这些来自大自然的启示，再加上人类的智慧，才有了人类现在如此发达的文明！这也正是“师法自然”的深刻含义。

从最早的原始社会，人类受两个石头撞击会产生火花的启发，学会了生火；受种子掉到地里就可以长出植物的启发，学会了耕种；到后来根据荷叶的启发，造出了雨伞；再到受鸟类依靠翅膀可以飞行的启发，发明了有翼飞机；根据鱼身体的“流线型”改良了舰船和潜艇。这些例子无不说明信息启迪作用的存在和影响。

1.2 信息技术

1.2.1 信息技术的含义

人们对信息技术的定义，因其使用的目的、范围、层次不同而有不同的表述。可以这样说，凡是能扩展人的信息器官功能的技术，都可以称为信息技术。因此，信息技术是人类开发和利用信息资源的所有手段和方法的总和，主要包括信息的产生、获取、变换、传递、存储、处理、显示、识别、提取、控制和使用技术等。

当前所说的信息技术实际上是一个新兴的技术群，主要是基于电子技术的信息技术（见图1.3），包括三个层面：基础信息技术、主体信息技术和应用信息技术。

基础信息技术主要包括微电子技术、光电子技术、真空电子技术、超导电子技术和分子电子技术等。信息技术和信息系统在性能上的提高，归根结底来源于基础信息技术的进步。

主体信息技术是指信息获取技术、信息传输技术、信息处理技术和信息控制技术等。这四项技术称为信息技术的“四基元”。

应用信息技术泛指由以上信息技术派生出来的针对各种应用目的的技术群类。它包含了信息技术在军事、工业、农业、交通运输、科学研究、文化教育、商业贸易、医疗卫生、体育运动、文学艺术、行政管理、社会服务、家庭娱乐等各个领域的应用，以及随之而形成的各行各业的信息系统。

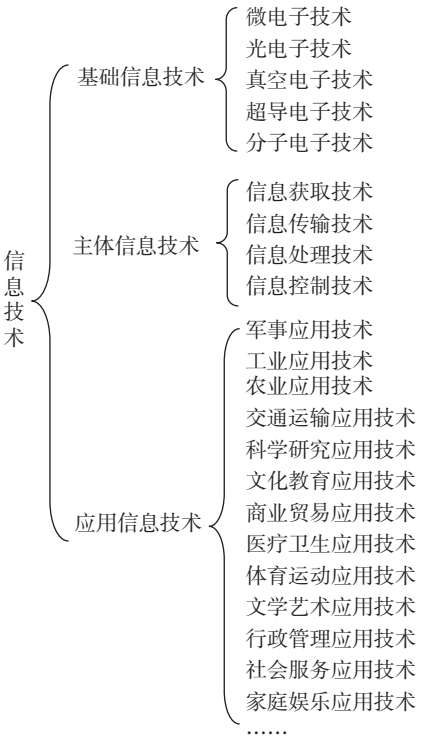


图1.3　信息技术群

当然，信息技术体系的层次划分只是相对的，而不是绝对的。例如，在主体技术与应用技术层次之间并不存在明显的界限。主体技术本身往往就是应用技术，比如一台计算机，它既是主体技术又是应用技术，通信系统的情形也是如此。又如，在主体技术与基础技术之间，虽然有着原则性的区别（主体技术一般是系统技术，直接扩展人的信息器官的功能，基础技术一般是标准部件或器件的制造技术，不能单独完成扩展人的信息功能的任务），但如果制造技术发展到如此地步，以致在制造过程中一次就能直接制造出一个完整的信息系统，而不只是标准的通用元器件，那么这种制造技术就已经属于主体技术，甚至是应用技术的范畴了。

1.2.2　信息技术的分类

信息技术可以按照多种方式分类。按表现形态的不同，信息技术可分为硬技术（物化技术）与软技术（非物化技术）。前者指各种信息设备及其功能，如显微镜、电话机、通信卫星、多媒体计算机。后者指关于信息获取与处理的各种知识、方法与技能，如语言文字技术、数据统计分析技术、规划决策技术及计算机软件技术等。

按工作流程中基本环节的不同，信息技术可分为信息获取技术、信息传递技术、信息存储技术、信息加工技术及信息标准化技术等。信息获取技术包括信息的搜索、感知、接收、过滤等，如显微镜、望远镜、气象卫星、温度计、钟表、Internet 搜索器中的技术等。信息传递技术指跨越空间共享信息的技术，又可分为不同类型。如单向传递与双向传递技术，单通道传递、多通道传递与广播传递技术。信息存储技术指跨越时间保存信息的技术，如印刷术、照相术、录音术、录像术、缩微术、磁盘术与光盘术等。信息加工技术是对信息进行描述、分类、排序、转换、浓缩、扩充、创新等的技术。信息加工技术的发展已有两次突破：从人脑信息加工到使用机械设备（如算盘、标尺等）进行信息加工，再发展为使用电子计算机与网络进行信息加工。信息标准化技术是使信息的获取、传递、存储、加工各环节有机衔接，提高信息交换共享能力的技术，如信息管理标准、字符编码标准、语言文字的规范化等。

日常用法中，有人根据使用的不同信息设备，把信息技术分为电话技术、电报技术、广播技术、电视技术、复印技术、缩微技术、卫星技术、计算机技术、网络技术等。也有人根据信息的不同传播模式，将信息技术分为传者信息处理技术、信息通道技术、受者信息处理技术、信息抗干扰技术等。

根据技术的不同功能层次，可将信息技术体系分为基础层次的信息技术（如新材料技术、新能源技术），支撑层次的信息技术（如机械技术、电子技术、激光技术、生物技术、空间技术等），主体层次的信息技术（如感测技术、通信技术、计算机技术、控制技术等），应用层次的信息技术（如文化教育、商业贸易、工农业生产、社会管理中用以提高效率和效益的各种自动化、智能化、信息化应用软件与设备等）。

1.2.3 信息技术发展趋势

信息技术在最近几十年里得到了空前的发展，速度惊人。但是信息技术仍然是非常年轻的技术，有着极大的发展空间。基于目前的现状，从信息技术总体的发展来看，今后将主要呈现出如下趋势。

1. 数字化

信息一旦具有数字形式，就容易加工处理、存储、传递和使用，有模拟电子技术所不具备的优势。数字信息的优点主要有：可处理性好、可压缩性强、无差错率高、可拓展性佳、保密性好，以及易于使处理设备通用化等，能够有效提高信息处理的速度和质量等。

信息的数字化意味着人类的工作、生活和从事其他社会活动都将数字化。数字化企业、数字化学校、数字化家庭、数字化图书馆、数字化社区、数字化城市、数字化地球……层出不穷，未来的社会就是“比特社会”，21世纪是“数字的世纪”。

2. 综合化

信息技术的综合是指各种不同信息技术的综合、各种信息业务的综合和各种信息网络的综合。各种信息技术在发展过程中实现综合，如计算机技术与通信技术的综合。就通信技术而言，无论传输、交换还是通信处理的功能，都采用数字技术，以实现网络技术一体化。由于充分发挥数字技术的优点，使信息网的总体效益实现最优。

3. 网络化

信息的快速传输和广泛共享依赖于发达的通信系统。通信线路的数字化、网络化是近年来一直进行的工作。现在已在公用电信网、电视广播网和计算机数据通信网三大网络体系的基础上，实现了三网合一，缩短了人们在时间和空间上的距离。

网络技术发展非常迅速。有专家指出，网络的宽带化、IP化将成为技术热点。网络宽带化包括基础网的宽带化和接入网的宽带化。IP化的发展趋势是一场技术革命。所谓IP化有两层含义：其一是指Internet上网；其二是指使用TCP/IP这种技术组建整个通信网，即所有的通信网设备，包括传输、交换、无线系统、各类终端，信令将都跑在统一的IP网络上。

4. 智能化

智能化是信息技术的又一个重要的发展趋势。智能分为生物智能、人工智能和计算智能。人工智能和计算智能的典型代表有智能网络、智能机器人、专家系统、智能计算机、智能终端，而人工神经网络则是对人脑的模仿，属于生物智能。智能网是近年来迅速发展的新的通信

技术，把交换机的交换逻辑与业务逻辑功能分开，分别由不同的网元来完成。智能网最终将实现电信网经营者和业务提供者能自行编程，使电信经营公司、业务提供者和用户三者均可参与业务生成过程，从而更经济、有效、全面地为用户提供各种电信业务。

信息技术的智能化趋势，将使因特网从一个单纯的大型数据中心发展成为一个更聪明的高智商网络，其中的个人网站复制功能将不断预估人们的信息需求和喜好。用户将通过该功能筛选网站，过滤掉无关的信息并将其以最佳格式展现出来。这种高智商网络将同今天的拨号音频电话一样方便，它与日常生活紧密结合的程度使人们甚至意识不到它的存在，人们可以使用电视、无线佩戴式麦克风等任意设备接入。

5. 多媒体化

多媒体技术是一种以计算机为核心，集图、文、声、像多种媒体的处理、传输、显示技术为一体的综合性技术。它是计算机、通信和大众传媒日益紧密结合的必然产物。多媒体化的主要表现是计算机的多媒体化、通信业务的多媒体化，以及各种多媒体应用的诞生和普及。

随着多媒体技术的发展，各种多媒体应用，如多媒体电子商务、网上购物、多媒体教学、远程医疗、多媒体电子出版物、视频点播等，似雨后春笋般涌现并快速走向普及。对于家庭来说，计算机、电话和电视最后将合为一体，组成家庭多媒体综合信息系统。

另外，信息技术还有并行化和集成化的趋势。并行化主要针对信息处理而言。集成化主要是指广泛利用系统集成和技术融合的办法来提高系统的性能，而且向着“创造性集成”方向发展。

总之，当前信息技术发展的总趋势是以互联网技术的发展和应用为中心，从典型的技术驱动发展模式向技术驱动与应用驱动相结合的模式转变。

1.3　军事信息和军事信息技术

1.3.1　军事信息

军事信息伴随着军事实践活动而产生。《孙子兵法》中的“知彼知己，百战不殆”，克劳塞维茨《战争论》中关于信息不确定性带来战争迷雾的著名论断，都强调了信息在战争中的重要作用。

军事信息是军事指挥、军事决策及执行决策所必需的各种情报、命令、消息、资料等数据的统称，它以数字、文字、符号、图表等形式，反映军事活动特征及其发展变化情况。在军事斗争和军事活动中，军事信息除了包括敌方信息、我方信息、战场环境信息等客观信息以外，还包括军人的思维信息等主观信息。其中，军人的思维信息是指军人在获得与军事有关的信息之后，经过自己的思维分析和加工处理后所形成的知识和观点。

军事信息除了军事领域的限定以外，其本质与一般信息并无根本区别。但军事信息由于其特殊的应用领域，我们应重点关注如下几个特性。

(1) 减少不确定性。美军在现代战争中特别强调“信息制胜”的思想，强调绝对信息优势，目的就是通过各种信息系统“知彼知己”，尤其是要消除或降低对敌方认知上的不确定性，以此消减战争中的迷雾。

(2) 可转移性。这一特性决定了军事信息能够被处理、融合和共享。同一内容的军事信息可以在不同的军事系统之间传递和复制。选择适当的载体，军事信息就可以在时间上和空间上实现转移。经过适当变换和处理后的军事信息可以被多次利用。

(3) 可压缩性。这一特性是指军事信息可以归纳、综合、概括，使其更加精练。通过对原始信息进行加工和浓缩，去粗取精、去伪存真，最大限度地减少其不确定性和多余部分，可以使军事信息增值，为指挥员决策提供更多帮助。

(4) 时效性。获得信息的时间不同，信息的价值就会不同。在指挥控制中，及时、准确、持续地获取军事信息，并保证信息在需要的时间到达需要的地方，可使己方遂行高速度、快节奏的作战行动，从而击败不具备这种能力的对手。

(5) 可度量性。一般来说，军事信息的质量可用完整性、正确性、及时性、准确性和一致性来衡量。而军事信息价值的度量则比较复杂，需要考虑多方面的因素，如军事信息系统效能、部队作战效能等。

无论是在冷兵器时代还是在热兵器时代，军事信息和军人的思维信息在战争中都发挥着重要作用。由于历史环境与客观条件的局限性，以前的信息获取能力、传输能力与处理能力比较落后，军人所获得的军事信息种类和数量都较少，经过加工处理后输出的思维信息也相应较少，因而它对武器能量释放的干预作用也就相对有限。由此导致当时人们对信息价值论的认识比较孤立和分散，并带有较大的主观随意性，智力与思维信息转换为武器能量释放的过程一直没有得到准确的描述与研究。随着现代信息技术在军事领域的广泛应用，特别是20世纪90年代以来几场局部战争发生以后，军事信息的作用越来越受到人们的重视，形成了争夺信息控制权和展开全面信息对抗的观点。

1.3.2 军事信息技术

军事信息技术是军事上用于信息获取、传输、处理、应用等技术的总称。它主要包括军事信息材料、器件、设备，以及系统的研究、设计、制造、综合集成和作战应用等方面的技术，是军事技术的重要组成部分。

军事信息技术种类多样。从组成的角度，军事信息技术可分为军事信息基础技术和军事信息装备技术两大类。军事信息基础技术是支持军事信息装备的微电子、光电子、真空电子技术，以及相关特种器件、电子材料、电源等的技术，是制造军事信息装备和信息化武器装备的核心。军事信息装备技术主要用于满足对军事信息的获取、传递、处理、控制和应用等各方面的信息需求。主要包括：指挥控制技术、预警探测技术、情报侦察技术、军事通信技术、导航定位技术、军用计算机技术、武器制导技术、信息对抗技术、信息安全技术、测量控制技术，以及军事电子信息系统技术等。另外，从信息流程的角度，军事信息技术还可分为信息获取技术、信息传输技术、信息存储技术、信息加工技术、信息应用技术和信息安全技术等。其中信息安全技术贯穿于整个信息流程。

由军事信息的种类可以看出，军事信息技术是综合性很强的技术，是国防技术群中的核心和骨干技术之一。信息技术用来迅速获取信息并快速处理和传送的特性，大大延伸了人的感官和触角，决定了其在军事上广泛的应用价值。军事信息技术所包含的学科内容非常丰富，已经形成门类齐全、技术复杂、特点突出的高技术群，主要表现在如下几个方面。

(1) 发展迅速。军事信息技术在现代军事技术群中是发展变化最为迅速的技术，如军事信息获取、传输和处理技术，导航定位技术，军用计算机技术等。从发明后就不断地更新换代，能力越来越强，水平越来越高。

(2) 应用广泛。军事信息技术已广泛、深入地应用于各类武器装备中，现代高技术战争的突出特征就是大量地使用信息化技术装备。

(3) 效果突出。例如，精确制导技术、导航定位技术应用于武器装备中，极大地提高了打击精度，使精确作战成为现实。

(4) 多学科交叉。军事信息技术综合性强，领域跨度大，学科分支多，如军事信息装备技术就包含了诸多的领域和学科。

(5) 渗透性、连通性强。使用军事信息技术可方便地把各种作战力量、作战单元、作战要素融合为一个结构合理的协调运行的整体。

自海湾战争以来的多次局部战争都呈现出信息化战争的趋势，军事信息技术作为国家技术实力的主要象征，得到各国的高度重视。军事信息技术为指战员在信息化战争中夺取信息优势、决策优势进而取得战场优势奠定了技术基础。主要表现在如下几个方面。

(1) 集信息获取、传输、处理和利用等各种技术之大成的指挥信息系统，为武装力量构造了“神经中枢”，成为军队战斗力的“倍增器”。

(2) 以电磁干扰、压制和网络对抗为核心，用以打击敌方信息系统的信息战，成为夺取信息优势和克敌制胜的关键。

(3) 信息化武器平台和以信息技术为核心的精确制导武器，成为军队的主战装备，并继续向智能化方向发展。

(4) 信息技术在军事上的广泛应用正在改变战争的形态，产生了“网络中心战”等新的作战理论，推动了军队的体制、编制和指挥方式的变革。

1.3.3　军事信息系统

信息系统是指以对信息进行收集、整理、转换、存储、传输、加工和利用为主要目的和特征的系统。信息系统具有多样性，即不同的信息系统具有不同的功能。但抽取其共性，信息系统是对信息进行采集、处理、存储、传输和管理，并向有关人员提供有用的辅助决策信息的系统。因此，无论何种信息系统，一般均具有六项基本功能：信息获取、信息处理、信息存储、信息传输、信息管理和辅助决策。

军事信息系统是应用于军事领域的一类特殊信息系统，指通过信息技术获取相关军事目标信息，并对信息进行处理和分发，为军队和武器装备的指挥控制及决策提供服务的综合信息系统。随着信息技术的发展，军事信息系统在战场空间的预警探测、侦察监视、军事通信、导航定位、指挥控制及综合保障等方面发挥着越来越重要的作用。

在形形色色的军事信息系统中，最具有代表性的一类是指挥自动化系统，它能够为军队作战、指挥、管理、保障等提供支持。在很多文献中又称其为“综合电子信息系统”。在指挥自动化系统的发展过程中，美军一直走在世界各国的前列。随着信息技术的发展，美军陆续提出了一系列具有指挥控制功能的军事信息系统，从C^2，C^3，C^3I，……，一直发展到目前美军正在研制的GIG（见图1.4）。各术语代表了不同年代、不同指挥控制功能的军事信息系统，详述如下。

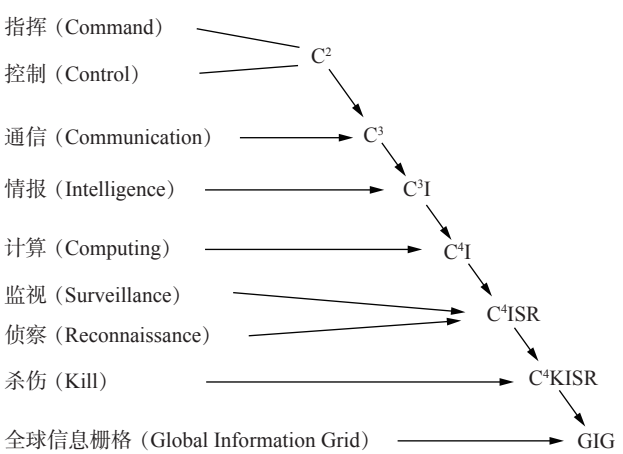


图1.4　军事信息系统术语演变示意图

- 20世纪50年代初期，随着电子技术的发展和在军事上的应用，美国首先提出了指挥和控制C^2系统；
- 20世纪60年代，由于通信技术的发展，通信作为新的要素被集成到指挥控制要素中，C^2系统扩展为C^3系统；
- 20世纪70年代末期，美国国防部又在C^3系统中加入了情报的概念，出现了C^3I系统；
- 20世纪80年代，随着超大规模集成电路的飞速发展，美国军用电子计算机发展迅速，广泛应用于军事装备中，计算机成为技术关键和指挥控制平台的核心处理设备，为C^3I系统提供了强大的信息处理支持，计算作为一个新的要素出现，于是把计算又加入了C^3I系统中，形成了C^4I系统；
- 20世纪90年代以后，情报、监视和侦察（ISR）在战场上发挥着至关重要的作用，形成了“传感器就是战斗力”和“发现即摧毁”的重要思想。随着信息技术内部各个分支学科之间的交叉渗透，C^4I和ISR集成为C^4ISR，实现了侦察预警与指挥控制的一体化；
- 2001年，为了使C^4ISR系统的各个要素与主战武器的杀伤更紧密地结合在一起，美国国防部先期研究计划局提出了C^4KISR概念，即将杀伤、摧毁能力嵌入C^4ISR系统中，通过将地、海、空、太空的各种传感器和指挥控制中心与武器平台集成为一体化网络，实现侦察/监视→决策→杀伤→战损评估过程等的一体化。
- 为了克服C^4ISR系统和C^4KISR系统各自独立开发、技术体制不统一、互联互通能力差的“烟囱式结构”等先天不足，美军于1999年首次提出了建立全球信息网格（Global Information Grid，GIG，很多文献中也将其称为“全球信息栅格”）的倡议，并于2000年3月，联合参谋部向国会正式提交了启动GIG项目的报告。GIG将把世界各地的美军指战员连接起来，在未来的信息化战争中，为他们提供联合作战所必需的数据、应用软件和通信能力，以获取信息优势和决策优势。GIG是未来战争能否从以武器平台为中心转向以网络为中心的关键，堪称“网络中心战”的“大脑”。

纵观上述以指挥自动化为代表的军事信息系统的发展史，不难发现，信息技术是军事信息系统产生、发展的前提和基础。正是美国信息技术的高速发展和具有的绝对优势，造就了美国先进的军事信息系统，使其得以在短短的50年里成为一个庞大的战争工具。

随着信息技术的发展，军事信息系统在战场空间的预警探测、侦察监视、军事通信、导航定位、指挥控制及综合保障等方面发挥着越来越重要的作用，已经成为军事作战指挥中不可或缺的重要组成部分。

1.4 军事信息技术与新军事变革

20世纪70年代以来，以高技术特别是飞速发展的信息技术为直接动力，工业时代的机械化军事形态开始向信息时代的信息化军事形态转变，国家安全战略与军事战略、国防建设与军队建设、战争与作战等领域都发生了深刻的变革，即新军事变革。新军事变革发生的根本原因是军事信息技术的迅猛发展，信息化是新军事变革的本质和核心，是各国军队转型建设的出发点和归宿。军事信息技术在新军事变革中扮演着关键角色，起着关键作用。

1.4.1　催生武器装备换代

武器装备是物化了的科学技术。军队和战争发展的历史阶段，往往是以武器装备的使用为标志进行划分的。从冷兵器时代、机械化战争时代到信息化战争时代，都明显打着科学技术的印记，都反映着新技术的累加增效过程。军事信息技术的高速发展，促使武器装备以日新月异的速度不断更新换代。

1. 信息化武器装备大量出现

(1) 智能化武器快速发展

从海湾战争到伊拉克战争的短短10余年内，以红外、激光、卫星定位等为代表的先进技术迅猛发展，使精确制导武器异军突起。在近几场局部战争中，精确制导武器战场使用率由8%迅速上升至90%以上。阿富汗战争中，美军整个作战体系经由C^4ISR联接实现了一体化。作战中，美军采用Link-16等数据链技术，将EQ-1“捕食者”无人机、RC-135V/M电子侦察机、U-2高空侦察机、E-8“联合星”飞机和RQ-4“全球鹰”无人机联接起来，实现了战场信息的互通与共享，从而提高了打击的灵活性和准确性，实现了真正意义上的“发现即摧毁”。计算表明，导弹战斗部爆炸威力若提高一倍，则杀伤力提高40%；而目标的精确识别和制导水平导致的命中率若提高一倍，则杀伤力提高400%。

(2) 电子战武器层出不穷

目前，多平台综合、多功能综合和多频段综合的电子战平台系统大量出现。红外光电对抗、反辐射武器等电子对抗系统的更新和发展，使战场电子对抗与反对抗趋于白热化，其“软杀伤”能力朝着宽频带、多目标、高能量、大功率和多平台方向发展。例如，电磁脉冲武器就是一种能在短时间内产生极强电磁波的信息战武器，它所释放的能量远远超过自然雷电释放的电磁脉冲，对信息化武器装备系统和信息网络具有巨大的破坏力。

(3) 新概念武器跃跃欲试

信息化武器装备体系的重要成员之一是新概念武器。新概念武器是指在工作原理、破坏机理和作战方式上与传统武器有很大不同，可大幅度提高作战效费比或形成新军事能力的高技术武器群体。目前，美、俄等国正在研制的新概念武器主要有定向能武器、动能武器、非致命武器三类。主要包括激光武器、微波武器、粒子束武器、动能拦截弹、电磁发射武器、动力系统熄火弹、次声武器等。值得提出的是，这些新概念武器的设计、研发与作战效能的发挥，都离不开计算机技术和网络技术的支持，而且计算机病毒武器、黑客进攻武器和思维控制武器等本身也是新概念武器的重要组成部分。目前，计算机病毒、黑客等技术已经较为成熟，预计将很快作为一种新概念武器装备用于实战。

(4) 天战武器咄咄逼人

组建航天部队，发展空间力量，运用天战武器攻击军用卫星、太空武器系统、弹道导弹和地球表面的重要战略目标，是当今世界各军事强国重点建设和发展的领域。目前，很多国家特别是美国和俄罗斯都在大力发展天战武器系统。正在研制或计划研制的天战武器主要有反卫星武器、反导武器和空天飞机等。以反卫星武器为例，它主要包括反卫星卫星、反卫星导弹、定向反卫星电子战武器、卫星干扰器等，可以攻击敌方部署在地球低轨道上的侦察、导航、气象卫星和航天飞机。美、俄的天战武器有望在2020年开始进入部署阶段，并形成实战能力。日本、印度等国也正在发展自己的反导武器，加快部署反导系统。

2. 军事信息技术成为武器装备作战效能的倍增器

信息化武器装备不仅提高武器装备本身的机械作战性能，更使目标搜索与识别、反应速度、对目标精确定位和打击等方面的能力得到全面提高。军事信息技术对于武器装备作战效能的倍增作用最为直接，不仅表现为同等兵力在战场上控制的空间范围倍数扩大，也表现为同等兵力的打击能力、摧毁能力、投送能力、打击精度、达成速率都成倍提高。详述如下。

(1) 信息化武器的打击能力大幅提升

军用信息技术的进步，可使武器装备具有更高的智能、更快的速度和更高的精度，可使指挥控制系统具有更精确、更安全、更快捷、更灵活的应变能力。利用信息所具有的力量控制和倍增功能，在较短时间内形成空前巨大的火力和毁伤力，突出表现在以下四个方面。

① 增大射程。信息化武器装备打破了地域性和时序性的限制，能够超越敌方的作战部署，直接打击敌方纵深目标。如美国的“战斧”巡航导弹射程为1850 km，俄罗斯Kh-101隐形巡航导弹的射程达到3000 km。增大射程可提高发射平台的生存能力，有利于遂行防区外作战。

② 提高精度。由信息为主导的精确打击技术、兵力火力的多维机动能力不断提高，火力运用方式由过去的大面积密集突击向精确打击转变。目前，美军精确制导武器的命中精度，近程弹药已达0.1~1 m，中程弹药小于10 m，远程弹药为10~50 m。

③ 提高反应速度。例如，在伊拉克战争中，参战的美军作战飞机都加装了“快速情报接收系统”，约占总数2/3的飞机在起飞前并未被赋予明确任务，起飞后先在战场上空待机，在空中接收实时的目标情报，而后发起攻击，从发现目标到实施打击的时间由过去的几十分钟缩短为几分钟。

④ 增大威力。武器精度越高，越趋于小型化，携弹量越多，作战效能越高。信息化武器更小、更轻、更灵巧，而杀伤力则丝毫不减。如采用GPS制导的“小型灵巧炸弹”，体积只等同于114 kg的炸弹，穿透力却相当于454 kg甚至900 kg的普通炸弹。由于体积小，重量轻，与普通炸弹相比，飞机可携带更多的“小型灵巧炸弹”，一次出动可攻击更多的目标。

(2) 老装备生存能力和作战性能全面激活

通过嵌入、融合信息技术或附加信息装置，能够提升武器装备的信息化含量，使其性能得到显著改善，功能有所加强，从而实现作战效能的跃升。利用“嵌入”法改造老装备，采取一代平台、多代负载，一种负载、多种平台等更新方式，可以降低投资风险和技术风险，节省1/3~1/2的费用，缩短一半以上的研制时间。

2005年，美国波音公司开始为B-52H轰炸机研制“战斗网络通信技术”系统。该系统可用于通信技术、战术数据链，提高B-52H轰炸机与其他军用系统各平台的信息共享能力，从而使这种老机型在高技术局部战争中重现活力。

在伊拉克战场，美国陆军选定了对维持战场优势具有重要意义的“阿帕奇”直升机、“艾布拉姆斯”坦克、M270多管火箭炮、“爱国者”防空导弹系统、“汉姆威”战车等17种装备，进行以信息化为主的综合改造。

俄罗斯、英国、法国等国也加强了老装备的信息化改造。如俄罗斯为“旋风”和“飓风”火箭炮及2519式自行榴弹炮加装了先进的“成就”火控系统，为它们配置“蜜蜂-1”无人侦察机，加快了侦察打击的一体化建设。英国和法国目前正在对其现役主炮AS90式155 mm自行榴弹炮和AUF-1式155 mm自行榴弹炮进行信息化改造，以增大射程和提高信息化作战能力。

(3) 武器装备的智能和自主性极大增强

军事高技术尤其是信息技术在军事领域的广泛运用，带来精确制导、遥感和探测、卫星通信和卫星预警、全球定位导航、隐身、激光、微光夜视、光电子等一系列高科技的迅速发展，

导致武器装备系统出现根本性甚至是断代性飞跃。传统武器多由物质和能量两大要素构成，而信息化武器系统的一个显著特点是追求物质、能量、信息三大要素的有机结合。正是这种结合，改变或部分改变了传统武器的纯粹的实体物质的机械性质，增加了除杀伤力和机动力之外的更重要的崭新能力——智能和自主性。

武器系统智能和自主性的提高，使作战效能获得了空前的提高。装有指挥系统的制导武器，是一种“会思考”的武器系统，能够自主搜索、发现、识别、攻击高价值目标，能够区分真假目标，能够筛选、判断和有选择地攻击敌方目标的薄弱环节和易损部位。智慧化无人作战系统，如无人机、无人坦克、无人潜艇等，能够自动进行态势评估和武器分配，可以在某些作战行动中替代人的进攻与防御。

3. 用信息技术构建武器装备的中枢神经

给武器装备装上“眼睛”和“耳朵”，只是武器装备信息化的基础，更关键的是用“神经”和“大脑”把各个部分联系起来，使其成为完整的信息化武器作战平台。通过联合、整合、嵌入、附加、链接等方法，将指挥、控制、通信、计算机、监视和情报侦察各要素加以系统集成，从而将探测器到武器装备操作人员的各个环节置于同一综合信息系统中。

(1) 为武器装备“植入”数字化装置

信息化的武器装备主要由“硬”和“软”两个部分组成。“硬”部分是指传统意义上的机械化武器装备，如坦克、步战车、舰艇、飞行器等；“软”部分则是指电子信息传感系统和指挥控制系统，它们具有战场感知、自动跟踪、信息处理、指挥控制等功能。利用信息技术，通过对传统的武器装备“植入”数字化装置，使其成为信息化的武器装备。

例如，美军对E-3预警机加装了如下四项数字化装置。

- 加装电子战情报支援系统，以探测、识别空中和海面的辐射源；
- 加装Link-16数据链的联合战术信息分发系统，以提供保密、抗干扰的数据分发通信手段；
- 增加计算机存储能力，以适应加装的电子战情报支援系统和Link-16数据链的联合战术信息分发系统，以及未来进一步扩充的需要；
- 加装惯性/GPS组合导航系统，提高导航定位精度。

这些措施使E-3预警机的作战效能得到大幅提高。

(2) 将武器装备“嵌入”作战平台

信息技术对武器装备系统的作用和影响，解决的不仅是单一武器的信息化问题，更重要的是通过数据链将武器装备“嵌入”作战平台，使相互独立的武器装备综合集成为新型武器装备系统，提升武器系统的整体效能和整体作战能力。

目前，世界各国都在加紧研制并部署综合作战系统，如美国的未来战斗系统、英国的未来快速奏效系统、法国的空地一体作战系统和俄罗斯的侦察打击一体化武器系统等。这些一体化武器系统，使地面、空中、海洋及天基的机动平台和武器、弹药之间出现强耦合关系，从传感器到射手，从单兵到作战单元，武器系统之间以及作战部队之间，实现信息快速流动，功能相互支持，形成远远大于各作战单元作战效能之和的总体作战效能。

(3) 使武器装备“进入”指挥控制网络

在信息化条件下，无论是武器装备还是作战平台，都不能“单骑突进”。信息技术应用于军事领域以后，最有代表性的概念是“系统集成”。

在信息化武器装备中，各军兵种仍可继续主宰各自传统的作战空间，有所不同的是以共同的软件、标准和规程，使各个武器装备平台实现互联、互通、互操作。

在整个军事系统中，按照作战职能建成侦察预警子系统、指挥控制子系统、精确打击与作战子系统、支援保障子系统。这四个子系统的功能紧密衔接，构成一个天网、地网一体化作战体系。它将整个战场上各军兵种的武器系统、作战平台、保障装备纳入统一的网络，使各类系统在更高层次、更大范围、更大规模上实现整体协调与效能优化。

1.4.2 触发战争形态变化

迄今为止，按物理形态划分，人类历史上完成了三次全面军事变革，即金属化军事变革、火药化军事变革、机械化军事变革。当今世界，第四次全面军事变革已经来临，即由机械化军事形态向信息化军事形态转变的新军事变革。

1. 信息化战争成为主要战争形态

战争形态是指由主战武器、军队编成、作战思想、作战方式等战争诸要素构成的战争总体面貌。其中，主战武器决定军队编成、作战思想和作战方式的变化，并由此产生不同的战争形态。当信息技术运用于军事领域并成为战场上的主导性力量时，就标志着人类社会经过徒手作战、冷兵器战争、热兵器战争、机械化战争后，进入信息化战争时代。

(1) 战争能量形态发生变化

信息日益成为重要的战争资源，目前成为引发战争能量形态变化的主导要素。人类战争从人体中心战、平台中心战进入网络中心战时代。以往战争中的“物质流”和“能量流”，只有依靠“信息流”的保障，才能更充分地转化为现实作战效能。信息控制作为独立的军事运动形式出现之后，人们在战争中能够有区别地、精确地运用力量，或者说信息技术使能量得以有控制地释放。这是信息化战争能量释放形态与以往战争能量释放形态的显著不同。

(2) 战争时空边界趋于模糊

信息网络覆盖战场各个领域，实现各类信息资源在不同时空的全维共享、实时交换，并使战场空间更为广阔透明，作战节奏和进程明显加快，作战样式呈现出非接触性、非线性的特征；战争边界由清晰趋于模糊，前线与后方、进攻与防御常常混为一体，难以区分。信息化条件下的战场，已经不是普通物理学意义上的自然空间，包括自然空间、网络空间、心理空间等。

(3) 战争对抗形式呈现新特征

信息化军队和信息化武器装备是基本作战力量，主要运用信息和信息手段，攻击敌方信息系统和思想信念，迫使敌方放弃对抗意志。制信息权的取得成为战争对抗的重中之重，围绕信息控制权的斗争广泛存在于战争的各个方面、各个层次、各个单元。非接触作战、非线性作战日益成为主要的作战手段，体系对抗越来越成为战场对抗的基本特征。

(4) 战争主体形态重新建构

从战争意志表达方式看，进行战争的主体将不再只是民族国家或国家集团、政党或团体，非国家主体、非政府组织、跨国公司、恐怖集团或“信息勇士”也同样能发动战争。从战争主体的价值理念看，信息化战争的胜负不再以歼灭敌有生力量多少为标准，而是强调从战略层面有选择地打击敌人军事目标，摧毁敌人的作战能力与抵抗意志。从战斗力建构方式看，由于信息技术和信息化武器装备的发展，不同军事力量将融为一体，合成军队的概念将被直接“固化”到武器系统中，武装力量将不再有军兵种之分，不同军事力量的作战功能或者说潜在价值是等效的。

2. 一体化联合作战成为基本作战形式

一体化联合作战是信息技术引起作战方式变革的必然结果，是未来信息化战场上的主要作战形式。与传统的机械化条件下的协同性联合作战相比，一体化联合作战基于信息网络系统，具有战场态势全维共享、作战力量系统集成、作战指挥实时精确等鲜明特征。

(1) 信息网络使战场态势全维共享

各种分散配置的侦察探测系统、指挥控制系统和火力打击系统集成为一个高效的网络体系，部队的信息获取、传递、使用、管理与共享的能力得到空前提高。信息网络将全维战场内的各类情报信息融为一体，形成共享的战场“通用态势图”，使从指挥官到单兵的所有作战单元、作战实体都能够实时感知全面的战场态势。

(2) 信息平台使作战力量系统实现集成

在信息网络的支撑下，参战的各军兵种、各类武器装备系统、各维战场空间和各种作战行动，形成一个结构紧密、反应灵敏，并能充分发挥各自优势的整体作战体系。以信息控制下的精确火力，打击敌作战系统中的关键节点，造成敌作战系统的结构性破损和功能性障碍；以精确、高效的指挥控制，使各单元、各子系统协调一致行动，形成整体合力，达成作战目的。这是系统与系统之间的对抗，是体系与体系之间的对抗，较量的是体系内各要素之间集成、联合的效率。

(3) 信息优势使作战指挥实时、精确

受技术手段的限制，机械化部队的指挥体系是纵长横窄的树状结构，指挥层次多，指挥控制容易滞后于战场实际，行动的精确性、应变性较差。一体化联合作战在信息技术支持下，采用“扁平网状”的指挥体系，指挥层次少，实时性、灵活性高。拥有信息优势的一方，对敌可实施信息垄断、信息威慑和信息攻击，对己可进行随机性、实时化、自主式协同。各种作战力量以目标和行动为中心，节点与节点、战斗单元与战斗单元之间联系紧密，信息交换迅速，能以更快的反应速度实施连续作战，打击敌方的关键目标和时敏目标，以及对不同目标实施同步打击。

1.4.3　推动军事理论创新

信息技术的发展直接推动了军事理论的创新，军事信息技术的广泛应用促使军事战略理论、军事作战理论及军队建设理论体系做出了适应性的改变。

1. 促进军事战略理论大发展

世界新军事变革和信息化战争，冲击着传统的军事战略理论体系。军事信息技术的迅猛发展，以及战争形态信息化和战争目的有限化的发展趋势，使得信息资源成为重要的军事战略资源，信息优势成为重要的军事战略优势，信息控制成为重要的军事战略选择。这使得信息化成为军事战略理论创新的主题词，“信息战略”、“信息威慑”、“信息保障”、“第五维战略空间”等新的战略指导理论，已经蕴藏于这一划时代军事变革之中。

2. 拓展军事作战理论新视野

随着信息技术发展及其在军事领域的广泛运用，信息化作战理论创新已成为军事理论创新的重中之重，“网络中心战”、“信息战役”、“信息行动”、“导航战”、“制敌机动”、“精确打击”、“全维防护”、“全频谱作战”、“非接触作战”和“非对称作战”等新的作战概念应运而生，为形成以信息化作战理论为核心的军事作战理论新体系奠定了基础。

3. 赋予军队建设理论高起点

信息化作战产生信息化需求，信息化需求牵引信息化建设。在当前新军事变革条件下，很多国家的军队已经提出不少军队信息化建设理论或原则，如“信息主导”、“系统集成”、“全能军队”、“一体化军队”等，核心是探索信息化战争的制胜之道。

1.4.4 引发体制编制改革

随着信息技术的发展，大量信息化武器配装到部队，对军队体制编制产生了直接而明显的影响。与信息化战争相匹配的军队体制编制，正朝着作战效能高、反应速度快、精兵合成好的方向发展。

1. 规模结构发生变化

随着机械化战争形态向信息化战争形态演变，军队的数量、质量与战斗力之间的关系将发生根本性变化，数量退居次要地位，质量跃居主导地位。于是，缩小军队整体规模、优化部队编制结构、发展精干军事力量，成为军队体制编制调整改革的主要目标。

2. 新型部队应运而生

在传统的机械化军事形态中，军队通常按主要作战领域、使命和武器装备，编有若干个军兵种。现代战场由传统的陆、海、空领域，逐步向陆、海、空、天、电、网等多维领域扩展，原有的某些军兵种可能改变或消失，而以信息技术为基础，与信息化战争息息相关的新型部队或特种部队，如信息战部队、太空战部队、机器人部队等应运而生。

3. 部队编成趋向综合

不同的战争形态需要不同组织结构的军队。军事信息技术发展的日新月异，使作战平台发生了根本性变化。信息化武器装备及合成军队的概念，将被直接固化到武器系统中。诸军兵种的界限日趋模糊，兵力、兵器合成将走向武器系统多种战斗效能的技术合成，这就为“扁平网状”指挥体制、“模块化”作战编组和一体化联合作战，提供了重要的物质基础。

1.4.5 促进教育训练转型

为了适应信息化武器装备的发展，满足信息化军队编制体制改革的要求，以及随着更多的信息技术直接应用于教育训练，机械化条件下的军事训练也在向信息化条件下的军事训练转变。

1. 一体化的教育训练体制

社会信息化和军事信息化进程的日益加快，使军事教育训练体制面临新的发展要求，必须树立起大系统、大合成、高效益、整体化的新观念，把军队院校的系统培训、部队基地化的岗位训练和地方科研机构的委托培养有机结合起来，逐步实现教育训练资源的集约管理和开放共享，提高教育训练保障效益，以一体化的联合教育训练，培养和造就高素质的新型军事人才。

2. 系统化的教育训练内容

以武器装备信息化、信息系统网络化、作战要素一体化为主要内容的新军事变革，快速改变着军队的面貌、军队的管理和运行方式。未来战争的形态，迫切要求军事教育训练的内容更新、层次更高、适应性更强。总的趋势是与军事信息技术发展同步，与战争形态变化合拍，与信息化武器装备接轨。

3. 超前化的教育训练模式

军事教育训练模式必然受教育训练目标、内容和条件的客观性制约。机械化战争时代的长时间经验积累式、尾随式等教育训练模式，已无法适应信息技术、武器装备和军事思想快速变化的进程，必须在抢占信息化人才制高点的总体战略和远景指导下，有针对性地实施具有前瞻性的军事教育训练，将不断发展的军事教育训练思想与军事信息技术、信息化武器装备紧密结合，把适用于机械化军队建设的教育训练模式转变为适用于信息化军队建设的联合培训模式、“理论牵引”模式、基地化训练模式，以及信息化的教育训练模式。

思考题

1. 什么是信息？它和消息、数据、信号、情报、知识等概念是什么关系？
2. 信息三要素包括什么？
3. 什么是信息技术，它是如何分类的？
4. 有一个信息源，发出0、1、2、3四种符号，且每种符号发出的概率相等。现在该信息源发出一串符号：2010132。计算这个符号串的信息量，以bit为单位。
5. 有一个信息源，发出0、1、2、3四种符号，每种符号发出的概率分别为$P(0)=1/2$，$P(1)=1/4$，$P(2)=1/8$，$P(3)=1/8$。计算该信息源的熵。
6. 信息技术对新军事变革有哪些影响？

第2章　信息获取技术

信息获取技术是应用信息科学原理和方法，实现并扩展人类感觉器官的功能，增强人类感知和认识事物的能力的技术。

信息的表达和传递依赖于物质和能量。一切信息获取技术所完成的功能，本质上都是先对携带着目标信息的介质进行采集，再转换为信息系统可识别和处理的形式，最后加以表示。例如，人们通过听觉获得信息，是因为空气的振动传递了声波，声波里携带着信息。人们通过视觉获得信息，则是因为可见光波起到了介质的作用。为了满足不同情况下对获取信息的不同要求，往往需要利用不同的物质作为介质，在采集、处理、表示等环节采用不同的方法。

按信息获取过程所利用的介质不同，信息获取技术可分为光电信息获取技术、雷达技术、声波信息获取技术、地面传感器技术和卫星导航定位技术。

2.1　电磁波基础知识

电磁波是天然的信息载体，在人类可预见的范围内，是唯一的理想信息传输介质。电子信息的产生、传播、探测、处理、对抗等行为，从根本上讲都是对电磁波进行操作，因此电磁波知识是学习信息技术必备的基础知识。

2.1.1　什么是电磁波

正如人们一直生活在空气中而眼睛却看不见空气一样，人们也看不见无处不在的电磁波。电磁波是电磁场的一种运动形态。电可以生成磁，磁也能带来电，变化的电场和变化的磁场构成了一个不可分离的统一的场，这就是电磁场。而变化的电磁场在空间的传播形成了电磁波。

电磁波具有波粒二象性：一是波动性，它具有与水波、声波、力学波一样的波动性，可以向周围或空间传播；二是粒子性，它占据空间，具有能量、动量和质量。电磁波有以下三个特点。

(1) 电磁波是高速运动着的物质（即具有能量、动量和质量），是物质世界的“长跑”冠军，它在真空中的传播速度为3×10^8 m/s。

(2) 电磁波没有静止的质点。

(3) 同一空间可以有无限多的电磁波同时存在，它“宽宏大度”而绝不排斥异己。

电磁波（如正弦波）具有振幅、频率、相位三个要素。振幅是振动的物理量可能达到的最大值，即波峰（或波谷）到横坐标轴的距离；频率是单位时间内完成振动的次数；相位用以表征波形上各点的相对位置，用角度的大小表示。

任何电磁波都有频率和波长两个基本参数。频率（f）的单位为赫兹（Hz）。频率的倒数称为周期，即振动一次所需的时间，单位为秒（s）。波长（λ）指波在一个振荡周期内传播的距

离，单位为米（m）。波长决定于振荡的频率和波的传播速度（c）。它们之间的关系是 $\lambda = c / f$。任何频率的电磁波，在真空中传播的速度均相同，即3×10^8 m/s，可视为一个常数，因此波长和频率成反比。也就是说，电磁波的波长越短，它的频率就越高；电磁波的波长越长，它的频率就越低。

2.1.2　电磁波波谱及波段划分

电磁波可分为两大类：无线电波和光波。它们有着共同的特点，如都是在空间传播的时变电磁场，都有极高的传播速度，只是波长（或频率）不同而已。

无线电波长在0.75 mm~100 km之间。按波长可划分为超长波、长波、中波、短波、超短波（米波）和微波（分米波、厘米波，毫米波）几个波段。按频率则可划分为与波段对应的甚低频（VLF）、低频（LF）、中频（MF）、高频（HF）、甚高频（VHF）、特高频（UHF）、超高频（SHF）和极高频（EHF）几个频段。无线电波的波段划分和应用如表2.1所示。

表2.1　无线电波的波段划分和应用

波段（频段）		波长	频率	应用范围
超长波（甚低频）		100~10 km	3~30 kHz	① 海岸–潜艇通信 ② 海上导航
长波（低频）		10~1 km	30~300 kHz	① 大气层内中等距离通信 ② 地下岩层通信 ③ 海上导航
中波（中频）		1000~100 m	300~3000 kHz	① 广播 ② 海上导航
短波（高频）		100~10 m	3~30 MHz	① 远距离短波通信 ② 短波广播
超短波（甚高频）		10~1 m	30~300 MHz	① 电离层散射通信（30~60 MHz） ② 流星余迹通信（30~100 MHz） ③ 人造电离层通信（30~144 MHz） ④ 对大气层内、外空间飞行体（飞机、导弹、卫星）的通信 ⑤ 电视、雷达、导航、移动通信
微波	分米波（特高频）	10~1 dm	300~3000 MHz	① 对流层散射通信（700~1000 MHz） ② 小容量（8~12路）微波接力通信（352~420 MHz） ③ 中容量（120路）微波接力通信（1700~2400 MHz）
	厘米波（超高频）	10~1 cm	3~30 GHz	① 大容量（2500路、6000路）微波接力通信（3600~4200 MHz、5850~8500 MHz） ② 数字通信 ③ 卫星通信 ④ 波导通信
	毫米波（极高频）	10~1 mm	30~300 GHz	再进入大气层时的通信

波长在1 mm以下的电磁波，统称为光波。在可见光之外，人们又先后发现了红外线、紫外线、X射线、γ射线等看不见的“光”。光波的波长由于比无线电波更短，通常使用微米（μm）和更小的单位“埃”（A），1 A = 10^{-10} m。光波的波段划分如表2.2所示。

表2.2 光波的波段划分

名称		波长
红外线	远红外线	25~1000 μm
	中红外线	1.5~25 μm
	近红外线	0.76~1.5 μm
可见光（红、橙、黄、绿、青、蓝、紫）		0.4~0.76 μm
紫外线		100 A~0.4 μm
X射线		0.01~100 A
γ射线		0.01 A以下

2.1.3 无线电波的传播特性与方式

电波既能在真空中传播，也能在介质中传播。电波从一种介质进入另一种介质时，会产生反射、折射、绕射和散射现象，同时速度也要发生变化；不同介质对一定频率的电波还具有吸收作用。电波的传播情况和电流不同，电流一般在导体中“流动”，而电波在理想导体中是不能传播的，金属壳体能够吸收电波，起到“屏蔽作用”；相反，电波在绝缘介质中容易传播。电波在传播过程中，由于能量的扩散和介质的吸收而逐渐减弱。离开波源越远，电波的强度越小。

根据介质及不同介质界面对电波传播产生的主要影响，无线电波传播方式分为地波传播、天波传播、视距传播、散射传播等（见图2.1）。

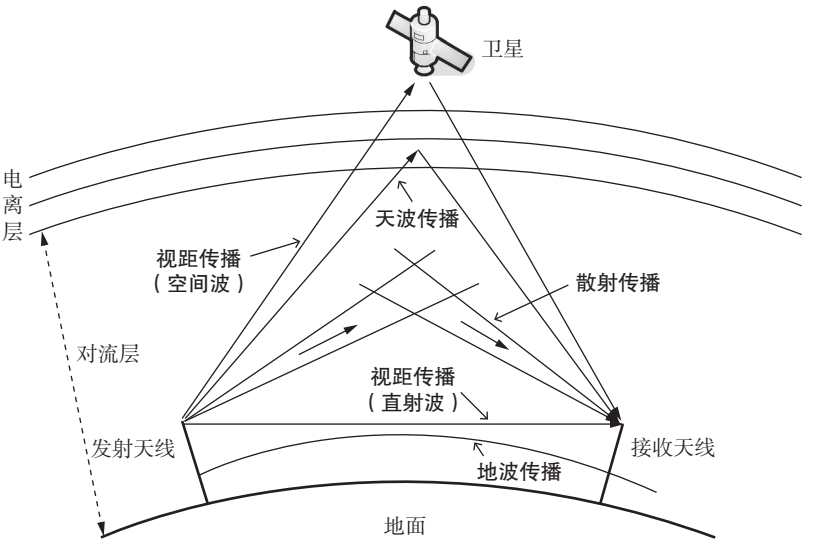


图2.1 无线电波的传播方式

1. 地波传播

无线电波沿着地球表面的传播称为地波传播。对于工作在低频段的系统，其天线低架于地面上，当其最大辐射方向指向地面时，电磁波主要以地波方式传播，实质上电波是绕着地面–空气的分界面传播的，这种传播方式适用于中、长波和超长波传播。

地波是沿着地表传播的，由于大地的电特性及地貌地物等因素不随时间很快地发生变化，并且基本上不受气候条件的影响，特别是无多径传输现象，因而地波传播信号稳定。另外，由

于中波、长波和超长波的波长很长，所以传输损耗小，作用距离远。地波传播主要的缺点是大气噪声电平高，工作频带窄。

地波传播主要应用于远距离无线电导航、具有标准频率和时间的信号的广播、对潜通信、地波超视距雷达等业务。

2. 天波传播

天波传播通常是指自发射天线发出的电波，在高空被电离层反射后到达接收点的传播方式。电离层是地球上空40~800 km高度的电离了的气体层，包含了大量的自由电子和离子。电离层既能反射电波，也能吸收电波，但对频率很高的电波吸收得很少。长波、中波和短波都可以利用天波传播方式进行工作，其中短波是利用电离层反射传播的最佳波段，可以借助电离层像镜子一样经多次反射，因而可以利用较小的功率进行远距离通信。在一年四季和一昼夜的不同时间，电离层都会变化，影响电波的反射，因此天波传播具有不稳定的特点。

3. 视距传播

电波在发射天线与接收天线之间可直视情况下的传播方式称为视距传播，是信号从发射天线到接收天线的直线传播。由于传播距离限于视线距离，此种工作方式要求天线具有强方向性并具有足够的架设高度。信号在传播中受到的主要影响是视距传播中的直射波与地面反射波之间的干涉。与利用电离层反射进行的“超视距”传播相比，视距传播特性稳定，外界干扰也较小。超短波和微波的传播特性如同光波，不能被电离层反射，在传播途径上没有或基本没有阻挡时，可忽略绕射，主要是视距传播，传播距离一般为10~50 km。为了增大传播距离，可采用微波接力通信，即在传送途中每隔一定距离设立接力站，像接力赛跑一样，把信息传到远处。

4. 散射传播

散射传播是利用对流层（离地面12~20 km高的大气层）或电离层中空气密度和离子密度的不均匀性，通过对电波的散射作用来实现超视距传播的。这种传播方式主要用于超短波和微波远距离通信，其中主要是对流层的散射通信。利用对流层不均匀体进行散射通信的频率，一般在100 MHz~10 GHz的频率范围内。100 MHz以下的电波散射效应很小，若频率过高，则大气吸收将显著增加。散射信号一般很弱，因此需使用大功率发射机、高灵敏度接收机和强方向性的天线。

2.2　光电信息获取技术

光电信息获取技术是以光波为介质，通过对目标反射或辐射的可见光、红外线或紫外线能量的感测，将其转换成电信号，从而获得目标信息的技术。该技术的隐蔽性好，战场生存能力强，分辨率高，抗干扰能力强，因此在军事上的应用十分普遍。光电信息获取技术主要包括可见光信息获取技术、红外信息获取技术、多光谱信息获取技术和紫外信息获取技术等。

2.2.1　可见光信息获取技术

2.2.1.1　基本原理

可见光本质上是波长在0.4~0.76 μm之间的电磁波。可见光获取技术就是通过截获目标反射的可见光来获取信息的技术。最常见的是，目标在太阳照射下发生了反射，用光学系统加以采集，把采集到的可见光作用于感光胶片，就能获得信息。或者把可见光信息经过光电转换转

化为电信号，再转换为目标信息。如果把获得的目标信息转换为电信号以后，不是把它制成相片，而是显示在屏幕上，并利用该技术对一个活动的目标连续地进行信息获取，并连续地显示图像，这就是电视摄像。夜间，由于只有青光、绿光等微弱的光线，因此需要通过微光夜视技术来有效地获取信息。

2.2.1.2 主要技术

1. 可见光照相

可见光照相是现代战争中广泛使用的一种侦察手段，能够获得最佳的地面分辨率，照片直观，易于判读，已大量用于空中（机载或置于气球上）、空间（星载）侦察与监视。美军在海湾战争、科索沃战争和伊拉克战争中都广泛利用了可见光照相技术，查明地形、地貌、地物和目标情况，测定目标坐标和评估射击效果等。

可见光照相侦察所运用的设备主要是可见光照相机，它是一种利用普通黑白和彩色胶片作为感光元件的照相机。根据其结构的不同，可分为画幅式、航线式和全景式三种。低空高速航空侦察机通常采用航线式和全景式相机，高空侦察通常采用画幅式或全景式相机。航天摄影通常采用多台不同功能的相机组成的相机系统，例如利用画幅式相机进行地物影像定位，利用全景式相机进行地面景物识别。

美国从1959年开始研制照相侦察卫星。1976年以前，卫星对地面拍照都用胶片。对胶片的处理有两种方式。一种方式是送回地面冲洗，这种方式分辨率较高，但时间上延迟很多；另一种方式是在卫星上自动冲洗，再以传真通信方式传回地面，这种方式减少了时间延迟，但分辨率较低。1976年，第五代照相侦察卫星“锁眼-11”（KH-11）入轨运行，采用了一种称为电荷耦合器件（Charge Coupled Device，CCD）的新型光电子器件，无须感光胶片，可直接获得与光学影像相对应的电信号，然后通过专用的数字通信系统，把图像信号传回地面。这样既保证了及时性，分辨率也达到了此前回收胶卷方式的水平。1989年以来，陆续入轨运行的第六代照相侦察卫星“锁眼-12”（KH-12），采用了先进的自适应光学成像技术，地面分辨率可达0.1 m左右。

2. 电视摄像

电视摄像的优点是能够实时地获得目标区域情况的活动图像。如我军陆军师编配了电视摄像机，用于侦察地面战场情况。外军的“锁眼-11”侦察卫星，以及许多类型的侦察机，也都装配了电视摄像机。

3. 微光夜视技术

在黑暗的环境中，有少量的自然光，如月光、星光、银河系的亮光和大气辉光（高层大气受太阳照射而发出的光），这些统称为夜天光。由于它们与太阳光、灯光相比十分微弱，所以又称为微光。微光夜视器材的基本工作原理是，先将目标反射的微弱夜天光信号转换成电信号，然后再将电信号放大，并用它去推动发光体发出可见光，即将电信号转换为人眼可见的光信号。目前军用微光夜视器材主要有微光夜视仪和微光电视两种。

微光夜视仪

微光夜视仪又称为星光瞄准具，诞生于20世纪60年代。它是利用对微光敏感的材料直接接收物体反射的夜天光并显示出来的装置。微光夜视仪主要由光学系统（物镜和目镜）、像增强管和高压供电装置组成。它的基本工作原理是：夜间景物反射的夜天光通过物镜进入像增强管，在像增强管内经过光电转换，得到增强，然后呈现在管内的荧光屏上，这样就能通过目镜

观察到被增强的夜间景物图像。微光夜视仪的关键部件是像增强管，它可以使图像的亮度增强5~10万倍。

微光夜视仪按用途可分为三类：(1) 微光观察仪，可用于夜间在前沿阵地对敌观察和监视，也可安装在侦察机、直升机上，用于侦察地面目标；(2) 微光驾驶仪，供坦克、车辆驾驶员戴在头盔上，黑夜中不开灯即可高速行驶；(3) 微光瞄准具，可装配于单兵武器及各种火炮上，进行夜间瞄准射击。

在马岛战争中，微光夜视仪发挥了显著作用。英军用直升机把特种部队运送到阿根廷军队后方，昼伏夜出，利用微光夜视仪进行各种侦察、袭击和破坏活动，并将重要的军事行动安排在夜间，凭借其夜视能力上的优势，使阿军不断受挫。海湾战争期间，美军步兵在地面战斗中也广泛使用了AN/PVS-7微光夜视眼镜。许多坦克配有夜间驾驶仪，车长配有观察仪，炮长配有瞄准具。

微光夜视仪的作用距离与环境照明条件及天气有关。在有星光、月光的条件下，可以观测到800 m距离处的人员和1500 m距离处的车辆。由于完全靠目标反射的夜天光工作，作用距离及观察效果受到气候和环境的很大制约，如树林或灌木丛的遮挡、雨雾等恶劣天气或阴云密布的全黑天，都会影响它的观测效果。

微光电视

微光电视实际是像增强技术和电视摄像技术相结合的产物，主要由摄像机、显示器和控制器三大部分组成。其中，摄像机里采用像增强技术的摄像管是微光电视的关键技术，它可以把像增强管荧光屏上的图像再转换成电信号，传输到远处并显示在电视屏幕上，还可供多人同时进行观察和实施战场监视、武器制导。微光电视具有比一般微光夜视仪更强的夜视效果和图像质量，在良好的天气条件下，微光电视摄像机的作用距离可达十几千米。

微光夜视器材的优点主要有三方面。(1) 图像清晰，对于1000 m距离之内的目标，其效果尤其良好。(2) 采用被动方式工作，即工作中本身不携带光源，因此不辐射任何电磁波，敌人难以发现。(3) 成本较低，因此曾是世界上应用最广、数量最多的军用夜视装备。

2.2.1.3　主要特点

各种可见光信息获取技术的共同优点是：(1) 图像直观，易于判读；(2) 高空对地面目标的分辨率高。其缺点是：(1) 难以发现隐蔽及伪装的目标；(2) 在照度太低及恶劣的气候条件下，信息获取能力大幅降低。正是与这两个缺点相对比，红外信息获取技术具有绝对优势，因此得到了很大的发展。

2.2.2　红外信息获取技术

2.2.2.1　基本原理

红外信息获取技术是利用目标反射或辐射的红外线来获取信息的技术。所谓红外线，是指波长为0.76~1000 μm的电磁波。首先，通过光学系统采集目标辐射或反射的红外线，使之作用于对红外线敏感的专门红外胶片，由此可以得到黑白的或假彩色的照片。也可以把红外信号通过光电变换转成电信号，再通过处理得到相片，或者在显示装置上显示出来，这些都称为红外成像技术。如果先向目标发出红外线，再采集目标反射的红外线，则称为主动式红外成像技术；如果采集的红外线是目标自身辐射的，则称为被动式红外成像技术。

2.2.2.2 主要技术

1. 主动式红外夜视仪（红外夜视仪）

红外夜视仪是一种用于夜间侦察、瞄准和驾驶的夜视仪器，诞生于20世纪40年代，并在第二次世界大战后期被美军和德军使用，现广泛地应用于部队侦察分队，包括红外观察仪、坦克和装甲车的红外驾驶仪等。

红外夜视仪由发射红外线的红外探照灯和接收目标反射红外线的夜视仪组成，通过红外探照灯去“照亮”目标，再利用目标反射回来的红外线成像来达到夜间观瞄的目的。在接通电源而不开启红外探照灯的情况下，可单独使用红外夜视仪来观测其他红外光源目标。

红外夜视仪具有成像清晰、成本低、使用维修方便等优点。但是，它的红外探照灯是一个很强的红外辐射源，在很远的距离上就会被敌方的红外探测装置轻易捕捉到，并招来火力袭击，成为被摧毁对象，这是它的致命弱点。1973年的中东战争中，埃以双方许多坦克都因装有这种夜视仪而遭击毁。所以近年来这种夜视仪逐渐被被动式红外夜视仪所取代，呈现被淘汰的趋势。

2. 被动式红外夜视仪（红外热像仪）

自然界中，温度高于绝对零度（−273.15℃）的一切物体，总是在不断地发射红外辐射。收集并探测这些辐射能，就能获得目标的有关信息。在此原理基础上发展起来的红外热像仪，是部队侦察和火力观瞄的主要夜视仪器。

红外热像仪能够利用光学系统，将目标发出的红外能量聚集在探测器上，探测目标各部分辐射的差异，获得图像细节，再通过可见光显示装置将其显示为可视图像。所显示的图像反映了目标与背景的温差信息，是“热图像”，故称为“热像仪”。红外热像仪还具有可探测到目标动态信息的特殊功能。根据图像中目标明暗的差别（即温度的高低），判断车辆或飞机等运动物体是正在发动还是刚刚停驶；还可以通过运动物体所留下的“热影子”，发现运动物体的行驶路线和原先的停放位置。

红外热像仪是被动探测设备，用于军事探测时不易被敌方发现，而且可以探测出隐藏的目标，穿透云雾的能力也比较强，因而在军事领域得到广泛应用。陆军将它用于夜间侦察瞄准、火炮及导弹火控系统、靶场跟踪测量系统；海军将它用于舰载火控、夜间导航及防空报警系统；空军将它用于飞机夜间导航、侦察及机载火控系统；星载热成像系统则可用于侦察地面、海上目标和对导弹预警等。作为光电侦察手段，红外热像仪的作用尤其突出。从可进行战场前沿侦察的单兵手持式热像仪，到车载、舰载、机载热像仪，都能在昼夜与不良天气条件下进行有效的工作。根据用途的不同，观测距离为2~20 km。

2.2.2.3 主要特点

红外信息获取技术的优点有以下几个方面。

（1）识别隐蔽及伪装目标的能力强。通常，对目标的隐蔽或伪装都是使目标对可见光不发生反射，或者使它对可见光的反射特性与周围物体相同或相似，因此这些目标在红外热像仪的观察下难免原形毕露。实践证明，将热像仪用于识别伪装，特别是发现隐藏在树林或草丛中的人员、车辆、火炮等，十分有效。

（2）能够透过“全黑”的夜暗，以及尘雾、雨雪等障碍，实现信息获取。

（3）作用距离远。可以达到电视摄像距离的两倍以上。将红外热像仪装在直升机上，可在1500 m高空发现地面的单兵活动；将其装在侦察机上，可在20 km高度发现地面的人群和行驶的车辆。

正是由于这些优点，红外信息获取技术特别是被动式红外信息获取技术，被认为是当前夜视技术最高水平的代表。

红外信息获取技术的缺点为：(1) 当目标与背景的温度差别较小时，图像较为模糊；(2) 在雨雪、浓雾严重条件下，其作用距离减小。

红外信息获取技术在识别隐蔽和伪装目标方面的成功运用，给了人们启发：既然同一物体对于不同波长电磁波的反射和辐射，一般都具有不同的特点，如果对此充分加以利用，应当可以更充分地获取更多的有用信息。正是基于这一思路，出现了多光谱信息获取技术。

2.2.3　多光谱信息获取技术

2.2.3.1　基本原理

多光谱信息获取（多光谱遥感）技术是在同一时间，对同一目标用多种不同的波段进行探测的技术。具体而言，就是将目标辐射或反射的各种电磁波划分成若干个窄的波段（光谱带），在同一时间内用几台遥感装置分别在各个不同光谱带上对同一目标进行照相或扫描，所得到的信息可以是图像形式的，也可以是数字形式的。对这些图像信息或数字信息进行加工处理，再与预先获得的各种目标辐射或反射的光谱信息进行对比，即可鉴别出目标的类型。

例如，绿色植物反射太阳的红外辐射的能力很强，但对于砍伐后用作伪装的植物，其反射红外辐射的能力则会大大减弱，而一般的绿色油漆对红外辐射的反射作用更弱。如果用几台遥感装置同时对目标分别拍摄红外、红色和绿色光谱带的照片，对它们进行处理叠放，则会形成“假彩色合成图像”。将其与真实彩色图像对比，就会看出：生长旺盛的植物呈红色，伪装用的植物呈灰蓝色，金属物体呈黑色。这样就能识别出经过伪装的目标。如果用多个遥感成像装置分别感测不同波长的红外辐射，经过对比和处理，识别效果就会更好。

采用此类技术，可以利用胶片得到照片，这种形式称为多光谱照相；对于采集的多光谱信息，可以用半导体敏感器件实现光电转换，再进一步处理，这种形式称为多光谱扫描；不同波段获得的图像信息也可以分别处理后在显示屏幕上相互叠加，获得合成的图像，如果将这一技术与电视技术结合，就形成了多光谱电视。

2.2.3.2　主要技术

多光谱信息获取技术是现代军队的有效侦察手段，但它的分辨率低于可见光，所以通常两者配合使用，互相补充，将两种装置同时装载在侦察飞机和侦察卫星平台上，执行战略、战术侦察任务。目前，常用的多光谱信息获取技术主要有多光谱照相、多光谱电视和多光谱扫描。

1. 多光谱照相

多光谱照相是多光谱信息获取技术中诞生最早的一种，是由普通航空照相机发展而来的。多光谱照相与普通照相的不同之处在于：普通照相接收的是可见光信息，而多光谱照相是在可见光的基础上向红外光和紫外光两个方向发展的，并通过各种滤光片或分光器与多种感光胶片组合，使其同时分别接收同一目标在不同较窄光谱带上所辐射或反射的信息。这样就可以得到目标的几张不同光谱带的照片。

多光谱照相技术主要有多镜头型、多相机型和光束分离型三种技术。其中，光束分离型照相技术的优点是结构简单、图像重叠精度高，但光束经过几次分光，对蓝色光的透射能量影响较大，降低了成像质量；多镜头型和多相机型照相机都存在着很难非常准确地对准同一地区、

重叠精度差、对成像质量也有影响的缺点，但多相机型多光谱照相机的灵活性较好，可适应多种需要，因而使用较多。

目前，用于侦察卫星的多光谱照相机，对地面景物的分辨率已达到5~10 m；机载的航空多光谱照相机的分辨率更高。

2. 多光谱电视

多光谱电视的工作原理与多镜头型及多相机型多光谱照相机相同，采用的也是滤光片分光方式。但它得到的是电视图像。在美国地球资源卫星上安装的是采用三台反束光导管电视摄像机的多光谱电视，它们分别拍摄蓝、绿、红三种颜色的地物图像，并将图像及时传回地面接收站。它与可见光电视摄像机相结合，具有很高的军事应用价值。目前，采用CCD的多光谱电视是重点发展方向。

3. 多光谱扫描

多光谱扫描是利用光学和机械扫描的方法接收地面目标景物辐射和反射的电磁波，通过多个分光片将这些电磁波按不同的波长分成若干波谱段（通道），并分别聚焦在敏感波长不同的半导体探测器件上，转换成电信号，用磁带记录下来或直接传输给地面接收站。

多光谱扫描与多光谱照相的根本区别是用半导体敏感探测器件代替了感光胶片。感光胶片只能感应可见光和近红外光，而多光谱扫描仪却可以覆盖从近紫外光、可见光、近红外光、中红外光到远红外光的大范围光波段。如砷化铟元件对1.1~1.8 μm波长的近红外光敏感，锑化铟元件对3~5 μm波长的中红外光敏感，等等。

多光谱扫描仪不仅工作波段的范围比多光谱照相机大大拓宽，而且它可把波段分得很窄、很多。目前，多光谱照相机可拍摄9个波段的照片，而多光谱扫描仪已提高到24个波段以上。也就是说，多光谱扫描仪把来自同一个目标的光波，分离成24个乃至更多个光谱波段进行记录，这就大大提高了识别伪装的能力。但是，多光谱扫描仪对地面目标的分辨率低于多光谱照相机，仅有20 m左右。

关于多光谱信息获取技术的应用，可以举出如下一些实例。

（1）美国“陆地卫星-5”。该卫星采用可见光至热红外光谱共7个谱段，分辨率为30~120 m，既可用于对大地进行资源考察、地图绘制等，也可用于对军事目标的侦察监视。海湾战争期间，该卫星每天飞越一次特定地区上空。美军曾利用其获得的图像绘制了科威特、伊拉克等国的地图。

（2）法国“斯波特-1”卫星。该卫星的多光谱模式使用3个谱段，分辨率为20 m。用途与美国“陆地卫星-5”基本相同，在海湾战争期间为美军提供了战区军事目标的大量遥感图片。

（3）我国“资源一号”卫星。1998年10月14日升空，2000年3月2日交付使用。其多光谱遥感所用谱段多于法国“斯波特-1”卫星，最高分辨率优于美国“陆地卫星-5”，达到了国际先进水平。稍后，两颗中国“资源二号”卫星又先后于2000年9月及2002年10月升空入轨，其总体性能和技术水平比“资源一号”卫星又有新的提高和发展。

2.2.3.3 主要特点

多光谱信息获取技术的主要优点是：（1）具有识别隐蔽及伪装目标的能力；（2）能够在夜间及不良气候条件下工作。多光谱信息获取技术的缺点主要是高空对地面目标分辨率不如可见光信息获取技术，此外其技术及装备也较为复杂。

2.2.4　紫外信息获取技术

2.2.4.1　基本原理

紫外线波长为0.01~0.4 μm，它在可见光的紫外端，故称为紫外线。从物理光学可知，在自然界中，太阳的紫外光大部分被大气中的氧和臭氧层强烈地吸收掉了，其中0.2~0.3 μm波长的紫外光几乎被完全吸收了，所以这段波长的紫外光区域（中紫外）通常称为“日盲”区；而0.3~0.4 μm波长的紫外光是能透过地球大气层最多的一段，因而该波段的紫外光区域（近紫外）通常称为“紫外窗口”。在大气层中，该段紫外光是均匀分布的。军事领域的应用主要利用中紫外的“日盲”和近“紫外窗口”特性来进行工作。

在“日盲”区，由于军事攻击目标的紫外辐射强于太阳的紫外辐射（如飞机的尾焰），所以目标很容易显现出来，相关的紫外系统正是利用“日盲”特性迅速而准确地探测和跟踪到攻击目标；在近紫外区，由于军事攻击目标挡住了大气散射的太阳紫外光，因而目标会形成一个“暗点”，呈现在均匀的紫外光背景上，相关探测和跟踪设备正是利用了这种“暗点”特性锁定了要探测和跟踪的目标。

紫外成像系统的组成与常规的微光成像系统相似，不同的只是把对微光灵敏的像增强器换成了对紫外灵敏的像增强器，再加上获取目标用的紫外镜头；而用于实时观察的目镜和保证像增强器能够正常工作的高压电源都与常规的微光成像系统中的一样。

2.2.4.2　主要技术

一般来讲，在军事上，紫外应用技术有成像和非成像技术之分，在有些场合并不需要图像，如导弹紫外制导等；而在另一些场合则需要对目标的紫外图像进行观察分析和处理应用，如紫外预警目标观察系统和紫外火控目标瞄准系统等。

紫外预警目标观察系统是综合预警系统中的一个子系统，它在综合预警系统必须保持电子静默和在雷达低空盲区的情况下，能够非常出色地完成预警任务。在太阳直射而使人的肉眼无法对特定目标进行观测时，它依然能很好地完成跟踪观察任务，尤其在相关系统遭到敌方红外干扰时，它还能作为红外的预备补充系统，完成红外系统的任务。

紫外火控目标瞄准系统是相关装备火控系统的子系统，它主要将紫外成像仪、激光测距仪、火控数据计算机等设备进行合理配置，以完成对目标进行探测、跟踪，并指挥火力装置打击敌对目标的任务。

2.3　雷达技术

雷达技术出现在20世纪30年代中期，主要是在第二次世界大战前夕，为了防空的需要而研制发展起来的，并在第二次世界大战中得到了广泛应用。雷达是英文“Radar”的音译，原意为“无线电探测与测距”，一般是指利用无线电波发现目标并测定其位置的设备。雷达能够发射某种特殊波形的无线电波（如微波和超短波等），接收并检测其回波信号的性质，以发现目标并提取出目标的速度、位置等信息。作为感知手段，雷达能够发现数百甚至数千千米的目标，可以不分昼夜地工作，从而为军队提供全天候预警能力。随着科技的进步，雷达的功能不断发展，性能不断提高，是各国大力发展的最重要的武器装备之一。

2.3.1 基本组成

雷达是一种有源装置，它用自身控制的电磁波照射来探测目标及其特性，不像光学照相机那样依赖于由非控制源反射回来的能量。雷达要发现目标和测量目标参数，就应该具备产生、发射、接收、测量和显示电磁波信号的设备，这些设备就是雷达的基本组成，如图2.2(a)所示。

(1) 发射机。其作用是在定时脉冲的控制下，产生并输出大功率高频振荡脉冲电流，送往天线。

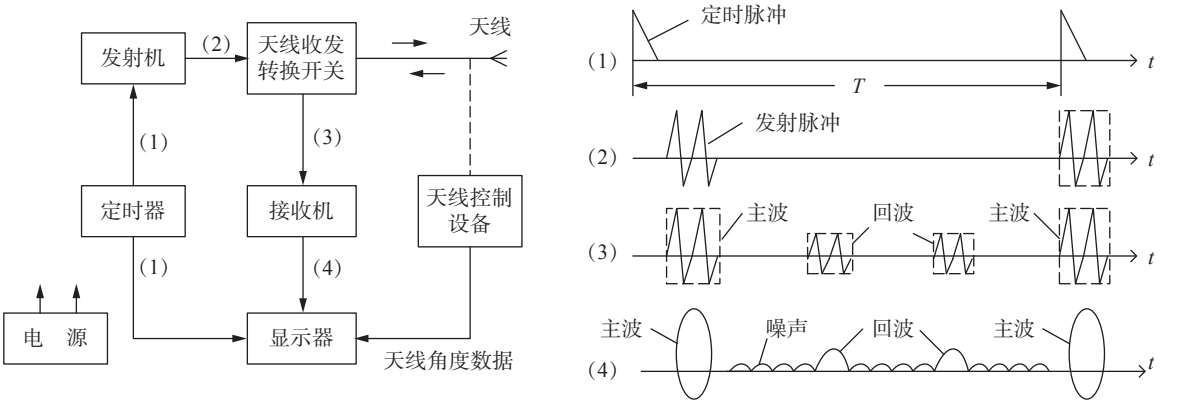
(2) 天线。其作用是将高频振荡脉冲电流转换成无线电波，并聚集成波束向空中发射，以及将目标反射的回波接收下来，送往接收机。

(3) 接收机。其作用是将目标回波进行处理后送给显示器，以便进行观察。

(4) 天线收发转换开关。一般情况下，雷达发射和接收使用的是同一副天线。发射时，天线收发转换开关使天线与发射机接通。发射一停止，收发开关便立即自动地将天线与发射机断开而与接收机连通，等待接收回波。

(5) 显示器。显示器是显示目标和测定目标位置的设备，分为距离显示器、方位显示器和高度显示器。

(6) 定时器。定时器是一个脉冲信号产生器，作用是使发射时间和显示器计时起点严格同步。为了达到这个目的，定时器同时向发射机和显示器发出一个时间基准信号，称为定时脉冲。在定时脉冲的触发下，发射机开始发射，显示器开始计时并在计时起点显示主波脉冲，如图2.2(b)所示。



(a) 脉冲雷达方框图　　(b) 脉冲雷达波形图

图2.2 脉冲雷达的基本构成

(7) 天线控制设备。其作用是控制和驱动天线旋转，同时不断把天线的方向数据送给方位显示器，一旦显示器上出现目标，就能读出目标的方位。

(8) 电源设备。保障供电。

雷达的工作过程描述如下。定时器不断产生连续的定时脉冲，用来触发发射机和显示器同时开始工作。发射机产生强功率高频振荡脉冲电流，经天线收发开关送到天线，天线将其转换成高频电磁波，并聚集成波束向空间发射出去。在天线控制设备的驱动下，天线转动并搜索目标。电磁波在空间传播的过程中，遇到目标时有一小部分电磁波反射回来，被雷达天线接收，称为回波信号。接收机将回波信号放大和变换后，送到显示器，显示器上就会出现一个回波信号或亮点，从而探测到目标的存在。

2.3.2　工作原理

1. 雷达定位原理

雷达最基本的任务是发现目标并测定其空间位置。测定目标时，通常采用球坐标系表示形式，即目标的斜距R（目标与雷达之间的直线距离）、方位角α（目标方向与正北方向的夹角）和仰角β（斜距与水平面之间的夹角），如图2.3所示。

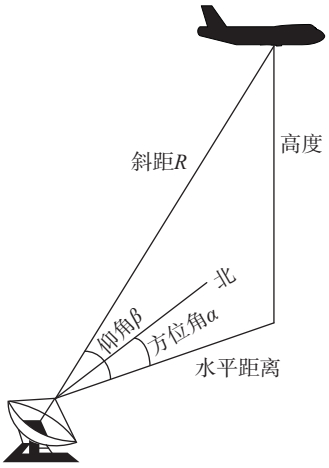


图2.3　雷达定位示意图

雷达波在空气中的传播速度是已知的，只要测出雷达波往返于目标与雷达之间所用的时间，就能按下面的公式计算出目标与雷达之间的距离（斜距）：

$$R = ct / 2 \tag{2.1}$$

其中，R为目标的斜距（单位为m）；c为电磁波传播速度，约等于3×10^8 m/s；t为电磁波往返于雷达和目标之间的时间（单位为s）。

实际上，显示器上的扫描线就是一种时间刻度。从扫描线起点到代表目标的亮点或波形之间的距离，正好对应于雷达机发出探测波到收到回波之间的时间间隔t。由于R与t一一对应，所以时间t直接反映了距离R。在具体的雷达机终端设备中，R数值的显示可以有多种形式。

雷达测角是利用天线波束的方向来实现的。雷达天线能将电磁波汇聚在窄波束内，辐射到所要求的方向上。当天线波束轴对准目标时，回波信号最强。根据接收回波最强时的天线波束指向，就能确定目标的方位角。水平方向移动的波束可以确定方位角α，垂直方向移动的波束可以确定仰角β。

目标的距离和方位角可以用方位显示器同时测量。图2.4为方位显示器的荧光屏图像。圆周上有方位刻度，圆心作为时间基线的起点，并代表雷达的位置，圆的半径为时间基线，并标有距离刻度。时间基线和天线同步旋转，当天线旋转一周时，荧光屏上的时间基线也旋转一周。雷达接收到的四面八方的目标回波都以辉光的亮点或短的圆弧形式显示在荧光屏的相应位置上，亮点中心所在的方位就是目标的方位，其数据可由距离刻度和方位刻度直接读出。

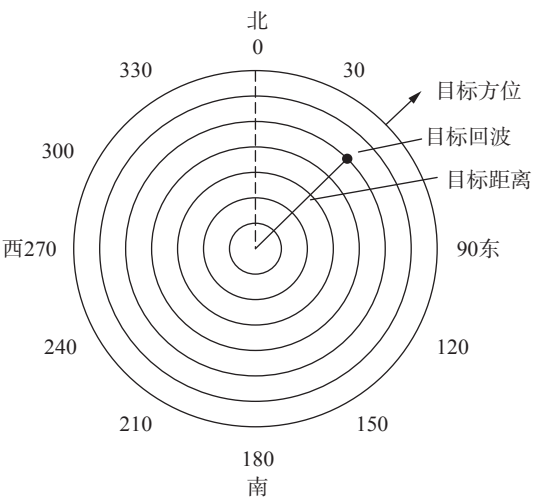


图2.4　方位显示器的荧光屏图像

2. 雷达测速原理

当火车向你开来时，声音不仅越来越响，而且音调越来越尖（即频率越来越高）。这种运动着的发声物体，由于具有速度而使声音频率发生改变的现象，称为多普勒效应。这种效应在电磁波和其他波动过程中也存在着。当波源与观察者有相对运动时，观察者接收到的频率与波源发出的频率不同。雷达就是利用多普勒效应来测量目标的运动速度的。

雷达发射信号的一般表达式为

$$s(t) = A\cos(\omega_0 t + \varphi_0) \tag{2.2}$$

其中，ω_0 为发射角频率，φ_0 为初相，A为振幅。雷达到目标的距离为R，则目标回波的延迟时间为

$$t_\tau = 2R / c \tag{2.3}$$

其中，c为光速。

雷达接收到的目标反射回波信号 $s_\tau(t)$ 为

$$s_\tau(t) = ks(t - t_\tau) = kA\cos\left[\omega_0(t - t_\tau) + \varphi_0\right] \tag{2.4}$$

其中，k为回波的衰减系数。

如果目标不动，则距离R为常数。如果目标与雷达之间有相对运动，则距离R随时间变化。设目标以匀速相对雷达运动，则在t时刻目标与雷达之间的距离$R(t)$为

$$R(t) = R_0 + v_\tau t \tag{2.5}$$

其中，R_0为$t = 0$时的距离，v_τ 为目标相对于雷达的径向运动速度，目标接近雷达时 v_τ 为负值，目标远离雷达时 v_τ 为正值。通常 v_τ 远小于光速c，所以延迟时间 t_τ 近似为

$$t_\tau = \frac{2R(t)}{c} = \frac{2}{c}(R_0 + v_\tau t) \tag{2.6}$$

回波信号与发射信号的相位差为

$$\varphi = -\omega_0 t_\tau = -\omega_0 \frac{2}{c}(R_0 + v_\tau t) = -2\pi\frac{2}{\lambda}(R_0 + v_\tau t) \tag{2.7}$$

它是时间t的函数，在径向速度 v_τ 为常数时，产生的频率差为

$$f_d = \frac{1}{2\pi}\frac{\mathrm{d}\varphi}{\mathrm{d}t} = -\frac{2v_\tau}{\lambda} \tag{2.8}$$

f_d 称为多普勒频率，正比于相对运动的速度而反比于工作波长。当目标飞向雷达时，多普勒频率为正值，接收信号频率高于发射信号频率；当目标背离雷达飞行时，多普勒频率为负值，接收信号频率低于发射信号频率。由此可以求得目标的运动速度。

3. 雷达识别目标原理

雷达在应用中有一个实际的问题是识别敌我。雷达发现的目标在显示器上以波形或亮点显示，难以辨出敌我，为此使用了敌我识别系统，所采用的技术装备称为二次雷达。即在雷达上添加询问器，它向目标发出一串编码指令。在目标上装有应答器，它收到询问信号并译码后，随即用同频率的密码回答。当询问器收到此回答后，如认为正确，即表示该目标为己方。

2.3.3 主要技术

1. 多普勒雷达技术

利用多普勒频移检测运动目标的雷达技术称为多普勒雷达技术。多普勒雷达用频率过滤的方法检测目标的多普勒频率谱线，滤除干扰杂波的谱线，即可从强杂波中分辨出目标信号。所以，多普勒雷达比普通雷达的抗杂波干扰能力强，能探测出隐蔽在背景中的活动目标，具有运动目标检测能力和显示能力。

多普勒雷达可以配置于飞机和卫星上，能检测并显示出地面人员和车辆的运动情况，不仅具有探测空中目标的能力，还增加了检测地面运动目标的下视能力，从而成为战场监视和侦察的重要感知手段。

2. 频率捷变技术

频率捷变技术是指发射的相邻脉冲的载频在一定频带内随机快速改变的脉冲雷达技术。它是雷达干扰与抗干扰斗争的产物，可以有效地对抗窄带瞄准式有源干扰，而且还具有加大探测距离、提高测角精度、抑制海浪杂波等优点。大多数军用雷达都采用这种体制，并已逐渐推广到民用船载雷达。频率捷变雷达可分为非相干频率捷变雷达和全相干频率捷变雷达两类，其中后者是主流。

3. 相控阵技术

相控阵雷达是指一类通过改变天线表面阵列所发出波束的合成方式来改变波束扫描方向的雷达。传统雷达若要改变波束的指向，只能由雷达天线的转动来实现。在那些具有庞大天线的雷达系统中，由于天线机械转动速度的限制，使得波束不可能在空间由一个方向快速地变到另一个方向，从而影响了雷达对多目标的探测与跟踪。相控阵雷达天线有别于机械扫描的雷达天线，可以减少或完全避免使用机械马达驱动雷达天线便可达到覆盖较大侦测范围的目的。

相控阵雷达的天线阵面由许多个辐射单元和接收单元（称为阵元）组成，阵元数目与雷达的功能有关，既可以有几百个，也可以高达几万个。这些阵元有规则地排列在平面上，构成阵列天线。每个阵元（或一组阵元）后面接有一个可控移相器，利用电磁波相干原理，通过控制各个移相器的相移量的方法来改变各阵元之间的相对馈电相位，从而改变天线阵面上的电磁波的相位分布。如图2.5所示，*MN*线上各阵元激发的电磁波的相位是同相的，而天线波束指向总是垂直于天线阵的“等相位面”，从而实现波束在空间按一定规律扫描，称为电扫描。

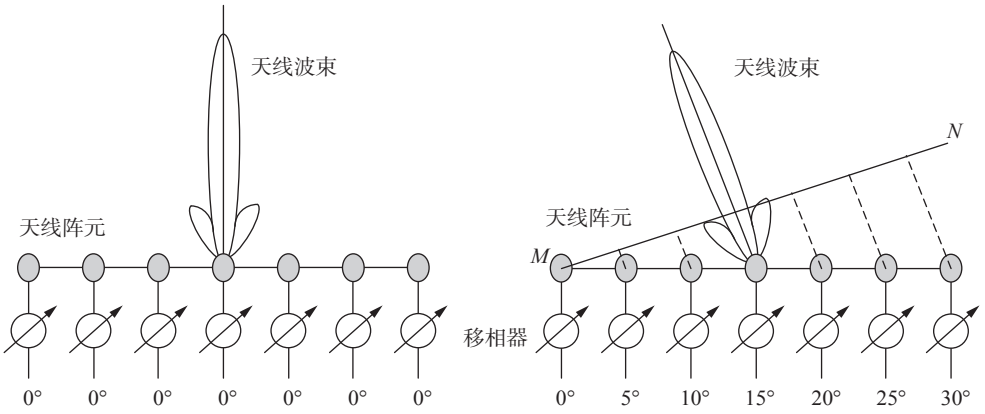


图2.5　相位扫描原理

阵元通过移相器可以被馈入不同相位的电流，从而在空间可辐射出不同方向性的波束。天线的阵元数目越多，则波束在空间中的可能的方位就越多。这种雷达的工作基础是相位可控的阵列天线，“相控阵”由此得名。由于摆脱了机械旋转，波束扫描的速度主要由移相器的相位变换速度决定，而该速度可达微秒量级，也就是说相位可以在一秒钟内变换几百次到上万次，因此天线波束的扫描速度极快，扫描的方式也非常灵活，从而能够达到多目标、高效益、高灵敏的参数效能。

相控阵雷达与常规机械扫描雷达相比，其优点如下。

(1) 波束指向灵活，能实现无惯性快速扫描，数据率高；

(2) 一个雷达可同时形成多个独立波束，分别实现搜索、识别、跟踪、制导、无源探测等多种功能；

(3) 目标容量大，可在空域内同时监视、跟踪数百个目标；

(4) 对复杂目标环境的适应能力强；

(5) 抗干扰性能好。全固态相控阵雷达的可靠性高，即使少量组件失效仍能正常工作。

目前典型的相控阵雷达共有两种组成形式：一种称为无源相控阵列，共用一个或几个发射机和接收机；另一种称为有源相控阵列，每个天线阵元有一个发射/接收装置，称为T/R组件，基本组成如图2.6所示。有源相控阵雷达与无源相控阵雷达的主要区别是，其射频功率的发射和接收通过阵元或子阵列中的发射/接收（T/R）组件实现。采用T/R组件，使有源相控阵雷达具有平均功率高、作用距离远、效率高、可靠性高等优点。

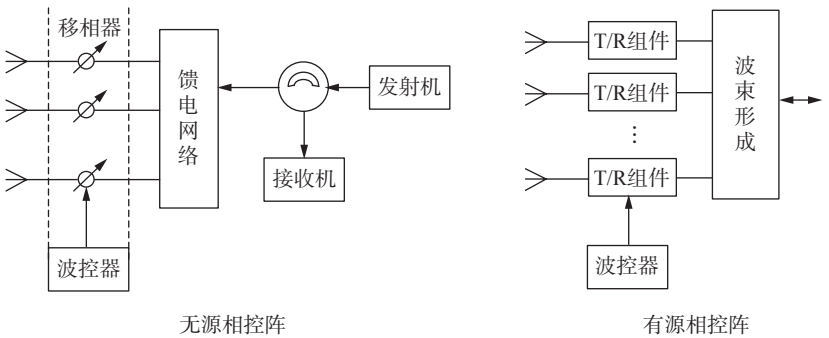


图2.6 相控阵雷达的基本组成

多功能相控阵雷达已广泛用于地面远程预警系统、机载和舰载防空系统、炮位测量、靶场测量等。美国“爱国者”防空系统的AN/MPQ-53雷达、舰载“宙斯盾”指挥控制系统中的雷达、B-1B轰炸机上的APQ-164雷达、俄罗斯C-300防空武器系统中的多功能雷达等，都是典型的相控阵雷达。随着微电子技术的发展，有源相控阵雷达得到了广泛应用，成为新一代的战术防空、监视、火控雷达。

4. 超视距技术

常规雷达利用了微波沿直线传播的特点来对目标定位，由于地球弯曲的表面造成了视距限制，常规的微波雷达无法观测到地平线以下的低空目标。超视距雷达利用短波的某些特性，来探测在地平线以下、海上或空中运动的目标。

超视距雷达按电磁波传播方式的不同，可分为天波超视距雷达和地波超视距雷达两类。天波超视距雷达利用大气的电离层对短波的反射特性，使电磁波在电离层和地面之间多次反射传播，接收目标的反射回波，从而能够发现目标，作用距离为1000~4000 km。地波超视距雷达利用了电磁波在地球表面的绕射效应，其工作频段为短波和超短波范围，作用距离较短，通常应用于海岸监视。

超视距雷达主要用于早期预警和战术警戒，是对地地导弹、部分轨道武器（包括低轨道卫星）和战略轰炸机的早期预警手段，预警时间可长达30分钟。在警戒低空入侵的飞机、巡航导弹和海面舰艇时，可以在200~400 km的距离内发现目标。与微波雷达相比，超视距雷达对飞机目标的预警时间约可增加10倍，对舰艇目标的预警时间可增加30~50倍。

5. 合成孔径与逆合成孔径成像雷达技术

雷达成像技术是20世纪50年代发展起来的一种新型雷达探测技术。在光学仪器中，孔径是指物镜的直径，它的大小决定透过光量的多少和分辨率的高低。雷达波是通过天线辐射和接收

的，天线就相当于光学仪器的物镜，孔径越大，辐射和接收的电磁波的能量就越大，雷达的作用距离就越远，分辨率就越高。由于受天线孔径的限制，常规雷达的横向分辨率远远低于光学和红外传感器的分辨率。为了克服天线孔径的限制，人们提出了“合成孔径”的概念，其思路是利用雷达与目标的相对运动，把雷达在不同位置接收的目标回波信号进行相干处理，实现天线的等效合成，使小孔径天线起到大孔径天线的作用，从而获得较高的目标方位分辨率。合成孔径雷达在波形上往往采用一些专门的信号处理技术来提高距离上的分辨率，从而最终提供高分辨率的雷达图像。到本书落笔时为止，最先进的星载合成孔径雷达的分辨率可以达到30 cm。

当目标不动而雷达平台运动时，这种雷达称为合成孔径雷达（Synthetic Aperture Radar，SAR）。当雷达平台不动而目标运动时，这种雷达称为逆合成孔径雷达（Inverse Synthetic Aperture Radar，ISAR）。如果将SAR和ISAR两种技术结合起来，便可在运动的雷达平台上，既对不运动目标成像，又对运动目标成像。

合成孔径雷达主要装在飞行器上，如飞机、卫星等。机载和星载合成孔径雷达具有观测面大、提供信息快、目标图像较清晰、能全天候工作、能从地面杂波中分辨出固定目标和运动目标、能有效识别伪装和穿透掩盖物等特点，广泛应用于战场侦察、资源勘测、地图测绘、海洋监视、环境遥感遥测等领域，是战场实时感知的最佳技术方法。

6. 激光雷达技术

激光雷达是由微波雷达发展而来的，基本原理与微波雷达相同，都是向目标发射探测信号，然后通过测量反射信号的到达时间、波束的指向、频率变化等参数来确定目标的距离、方位和速度。只是激光雷达利用激光束来工作，波长比微波短得多，只有0.4~0.75 μm。它不仅可精确测距，而且能精确测速、精确跟踪。

与微波雷达相比，激光雷达把辐射源的频率提高到光频段，比毫米波高出2~4个数量级。这使之能探测迄今所遇到的任意微小的自然目标，包括极细的导线和发射的粒子。通过探测激光回波的多普勒频移，可感知目标的小量振动。由于这种频移正比于载波频率，故激光雷达探测微小振动的灵敏度比毫米波雷达高两三个数量级。这对目标识别非常有利。

激光雷达的带宽比毫米波雷达小两三个数量级，加之它的高空间分辨力及高灵敏度，使我们可以获取目标尺寸、形状、速度、振动及旋转速度等多种信息，实现对其准确识别和跟踪。激光优异的单色性和极小的脉冲宽度，使激光雷达能排除背景和地面杂波干扰，减小噪声影响，因而能探测超低空目标，可用于跟踪发射初始段的导弹和巡航导弹。例如，洲际导弹若释放大量电磁假目标，或借助小型核爆炸构成人为的反射微波电离层，就足以使微波雷达失效，但对激光雷达的影响却很小。

激光雷达的天线尺寸比微波雷达的小得多。例如，同样从地球照射月球上1 km^2的区域，激光雷达的发射天线（光学望远镜）直径约为30 cm，而微波雷达天线的直径则约几千米。可见激光雷达更适于车载、机载和星载用途。

激光雷达的缺点是：由于大气对激光的散射和吸收比微波严重，尤其是有云、雾、雨时，激光雷达的作用距离小。另外，由于激光散角小，大面积搜索时容易丢失目标，故不宜作为搜索雷达使用。它与微波雷达结合使用，可扬长避短。

2.4 声波信息获取技术

声波信息获取技术是以声学原理为基础，根据被探测目标在声波传播介质（如地层、大气和水等）中发出的声频振动，利用电子装置处理获取的声波信息，以实现对目标的识别和定位的技术。以空气或地层中的声频机械波为介质的声测器材和振动传感器将在下一节介绍，本节主要讨论水下声波信息获取技术。

以电磁波为介质是获得信息的重要途径，但在水中，电磁波的能量会被强烈地吸收。即使是具有强烈光柱的探照灯，一到水下最多只能照射几十米，而雷达波一入大海就会很快被海水吸收而产生热量损耗掉，几乎是寸步难行。而声波就大不相同了。它在水中的传播速度接近1500 m/s，几乎是它在空气中传播速度的5倍。声波在空气中损耗很快，而在海水中损耗很小，传得很远。人们利用声波在水中良好的传播特性来获取水下目标信息，应用最广泛、最重要的一种装置就是声呐。

2.4.1 声呐的任务和分类

声呐，是英文缩写“SONAR”的音译，原意为“声音导航与测距”（Sound Navigation and Ranging），是利用水中声波进行探测、定位和通信的电子设备。其任务包括对水中目标进行搜索、警戒、识别、跟踪、监视和运动参数的测定，以及进行水下通信和导航等。在军事侦察中，声呐主要用于为水下潜艇和水面舰船及反潜飞机实施反潜战、反水雷、反水面舰艇战斗提供情报保障，同时用于海岸、港口的防御警戒，御敌入侵等。其作用是探测、分类、确认和定位跟踪海洋目标，包括水雷和鱼雷等。自从潜艇问世以来，以水声技术为基础发展起来的声呐，就成为反潜探测的重要工具和手段，特别是弹道导弹和巡航导弹及核潜艇的出现，使声呐的作用更为突出。

声呐的分类方式有很多种。按工作方式，可分为主动声呐和被动声呐；按装备场合，可分为水面舰艇声呐、潜艇声呐、航空声呐及海岸声呐等；按战术用途，可分为搜索警戒声呐、识别声呐、探雷声呐等；按基阵携带方式，又可分为舰壳声呐、拖曳声呐、吊放声呐、浮标声呐等。这里主要讨论主动式和被动式两类声呐技术。前者通过向目标发出声波，再接收其回波并加以处理，从而获取目标信息。后者不向目标发出声波，仅仅接收目标自身在运动中发出的声波来获取目标信息。

2.4.2 主动声呐

2.4.2.1 基本组成与工作原理

主动声呐又称有源声呐，通过水声换能器及其组成的阵列，将探测电信号转换为一定频率的声波，形成水下声信号。当此信号遇到目标时，形成反射回波，声呐将其接收并转换为电信号，经过处理，就可发现水下目标的存在，并可将目标定位。

主动声呐的基本组成如图2.7所示，由换能器基阵、收发开关、发射机、接收机和信号处理（包括信号调理、模数转换、波束形成、复数基带解调和快速傅里叶变换等）设备，以及系统监控显示设备等组成。现结合该图简述其工作原理。

图2.7所示的基阵，是由若干水声换能器以一定几何形状排列组合而成的阵列。水声换能器是发射和接收水下声信号的装置，其中应用最广泛的是完成电声转换的水声换能器，包括把电信号转换为水中声信号的水声发射器，以及把水中声信号转换为电信号的声波接收器（水听器）。

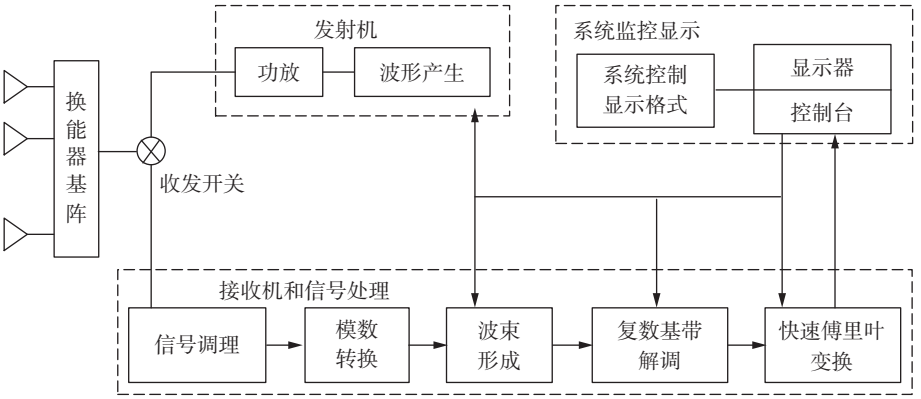


图2.7　主动声呐的基本组成

水声换能器有可逆式与不可逆式之分，前者既可用作水声发射器，又可用作水听器；后者则不能兼有此两种作用。广泛采用的压电水声换能器和磁致伸缩式水声换能器都属于前者。水声换能器虽然也可以单独使用，但多数情况下是以组成阵列的形式工作的。水声换能器基阵可以是单独的发射基阵或接收基阵，也可以是收发合一基阵。由上述可见，声呐的水声换能器及其组成的基阵，其功能和工作方式十分类似于雷达的天线，只不过前者收发的是水声波，后者收发的是电磁波。

主动声呐大多采用脉冲波体制，工作频率从较早期采用的超声波频段（20~30 kHz）向低频段发展，目前一般为3~5 kHz；发射功率可达100~150 kW，最高已达1 MW。

主动声呐的工作过程如下：控制分系统定时触发发射机的信号发生器，产生脉冲信号，经波束形成矩阵和多路功率放大，再经收发转换网络，输入发射基阵，形成单个或多个具有一定扇面的指向性波束，向水中辐射声脉冲信号，也可能辐射无方向性的水波声脉冲信号。在发射基阵向水中辐射声脉冲信号的同时，有部分信号能量被耦合到接收机，作为计时起点信号，也就是距离零点信号。声呐发射的水波声脉冲信号遇到目标就形成反射回波，回传到声呐的接收基阵，被转换为电信号，经过放大、滤波等处理，形成单个或多个指向性接收波束，在背景噪声中提取有用信号。基于与雷达定位类似的原理，可以测定水中目标的距离和方位。测得的目标有关信息最后在终端设备输出。终端设备可以是显示器、耳机、扬声器、记录器等。

2.4.2.2　功能特点和应用

主动声呐的主要优点为：（1）可以探测静止无声的目标；（2）既可测定目标的方位，又可测定目标的距离。20世纪90年代初期的主动声呐探测距离可达十至数十海里[①]；利用数字多波束和单波束电子扫描技术，可以实现对目标的水平全向或三维空间搜索，可以搜索和跟踪多个目标。主动声呐的主要缺点是隐蔽性差，增加了受敌干扰和攻击的可能。

主动声呐是水面舰艇声呐的主要体制，完成对水下目标（如冰山、暗礁、沉船、海深、鱼群、水雷和关闭了发动机的隐蔽潜艇等）的探测定位等任务。例如，法国大型水面舰艇装备的SS-48型声呐，在具有良好水文条件，目标为中型潜艇且本舰航速为19~20节的情况下，其主动探测距离在全向发射时为15海里，定向发射时为20海里。

① 1海里＝1852 m——编者注。

2.4.3 被动声呐

2.4.3.1 基本组成与工作原理

被动声呐又称无源声呐或噪声声呐，本身不发射声波，依靠接收目标辐射的声波作为目标检测和对之估值的基础。此目标辐射的声波一般是目标自身发出的噪声，包括螺旋桨转动噪声、流体动力噪声和发动机机械振动引起的辐射噪声等，或者是目标声呐的辐射声波。被动声呐以此目标噪声作为信号，通过换能器基阵接收此微弱信号，通过空域、时域、频域信号处理，提取并检测有用信号，分析并估算回波的方向、距离，进行跟踪并判断信源的性质和类型等。

被动声呐的基本组成如图2.8所示，由接收换能器基阵、信号调理（放大、滤波和动态压缩）、模数转换、波束形成、数字滤波、检测/估值统计处理器和监控显示等功能部件组成。

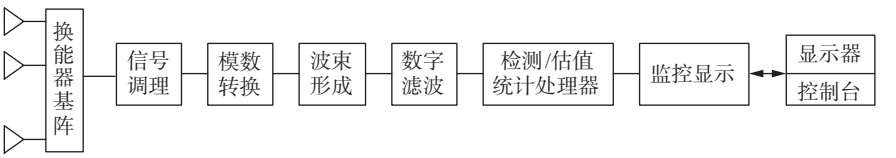


图2.8 被动声呐的基本组成

对比图2.7与图2.8可发现：被动声呐组成与主动声呐的接收部分基本相同，但要求更高，因为被动声呐接收信号为负信噪比，需采用更多、更严格的处理措施，以获取更高的处理增益。

2.4.3.2 功能特点和应用

被动声呐的优缺点恰与主动声呐相反。它不能探测静止无声的目标；一般只能测定目标的方位，不能测距。最主要的优点是隐蔽性好。因此潜艇声呐在大多数情况下都以被动方式工作，对水中目标进行警戒、探测、跟踪和识别。被动声呐系统在鱼雷、水雷中也是最主要的传感器之一，它的这种隐蔽的、相对低功率的方式具有较大的覆盖面，且能够有效地启动静止的鱼雷，启动并操作移动式武器。海岸声呐工作通常也以被动方式为主。

2.5 地面传感器技术

地面传感器是20世纪60年代出现的一种辅助性战术侦察器材，通常是指能对地面目标运动所引起的电、磁、声、地面振动和红外辐射等物理量的变化进行探测，并将其转换成电信号的设备，主要用于执行预警、目标搜索与监视等任务。地面传感器受地形、地物和气候的限制很小，并具有实时、隐蔽和不间断地自动侦察与监视的功能，可布放在战场侦察雷达、光学器材、夜视器材的“视线”达不到的山地或丛林地区。人员、装备在地面上运动时必然发出声响，引起地面振动，或使红外辐射发生变化。在一定条件下，携带武器的人员和装备还会引起电场、磁场的变化。地面传感器正是通过探测这些物理量的变化来发现与识别运动目标的。

由于地面传感器能够有效地弥补雷达和光学侦察系统的不足，扩大了战场信息探测的时空范围，因此许多国家的军队都十分重视对它的研究与应用。

2.5.1 基本原理

地面传感器通常由探测器、信号处理电路、发射机和电源四部分组成（见图2.9）。

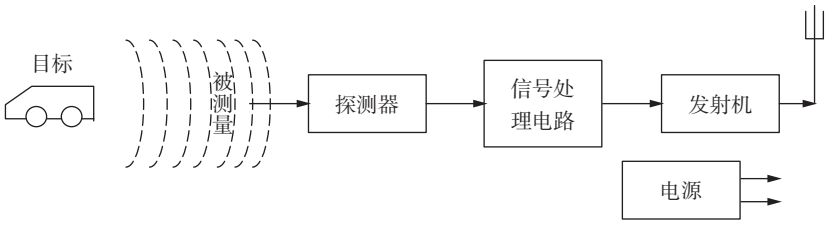


图2.9　地面传感器的基本构成

不同类型的地面传感器的主要区别是探测器不同，即接收的被测量不同，因而探测原理也不相同。其余部分则基本相同或互相通用。地面传感器的工作过程是：运动目标所产生的地面振动波、声响、红外辐射、电磁或磁能等被测量，由探测器接收并转换成电信号，由信号处理电路放大和处理，送入发射机进行调制后发射出去。

2.5.2　主要技术

1. 振动传感器

振动传感器是普遍使用的采用微型信息处理器的地面传感器。主要是通过仪器的振动探头（拾振器）来捕捉人员或车辆活动引起的振动信号，以发现远距离处运动的敌方人员和车辆。战场上设置振动传感器时，通常将拾振器设在地表层，并严密伪装。当人员和车辆经过附近时，传感器便将目标引起的地面振动信号转化为电信号，经放大处理后发给监控中心，从而进行实时的战场监视。如美军远程巡逻队使用的一种轻型振动传感器，能在5分钟内设置完毕，并可在1800 m的距离以内进行监控。

振动传感器的主要优点如下。

（1）灵敏度高，可探测到地面微小的振动信号；

（2）探测距离远，可探测到距传感器30 m以内的人员和300 m以内的车辆；

（3）耗能少，传感器内的一节电池可连续工作数日；

（4）布设手段多，既可人工安放，又可机载空投，还可火炮发射；

（5）具有一定的目标分辨力，可有效地辨别人为振动或自然扰动，并能准确地区分人员和车辆。

其主要缺点如下。

（1）在松软土质条件下，探测范围明显减小；

（2）沟壕水溪等可完全阻断振动信号的传播，在沼泽、滩涂、水网等地区，传感器则无法工作；

（3）分辨率还不是很高，若要更准确地鉴别目标，如判断是徒手人员还是武装人员，是履带车辆还是轮式车辆，目前振动传感器还做不到，所以在战场上往往与其他类型的传感器联合使用。

2. 声响传感器

声响传感器主要利用能够灵敏地接受战场声响的探测装置，分辨侦察区域内敌方目标的性质，监视其活动情况。其探测器是一个传声器，俗称话筒。它将目标的声音信号转换成电信号，发送给监控中心，再还原为声音信号。它对目标的分辨能力极高。如果探测的目标是人员，不仅可以探测到目标的数量和运动情况，还可以判明其国籍；如果是车辆，则能够判明车辆的种类。特别是能够准确地分辨出是人为的声响还是自然声响。声响传感器对人员正常谈话

声的探测距离可达40 m，对运动车辆的探测距离可达数百米。如美国陆军使用的一种可悬挂在树上的“音响浮标”，探测距离高达300~400 m，已接近人的听力范围。

声响传感器的主要缺点是耗电量大。因此，主要是采用人工指令遥控其启动或关机，并与振动传感器联用。平时振动传感器处于工作状态，声响传感器则处于关机状态。振动传感器探测到目标后再启动声响传感器，把振动传感器耗电量少与声响传感器鉴别目标能力强的优点结合起来，达到取长补短、相互配合完成探测任务的目的。

3. 磁性传感器

磁性传感器的探测器为一个磁性探头。探头工作时，能连续发出无线电信号，并在其周围形成一个磁场。当金属物体进入这个磁场后，磁场的信号就会受到扰动。由于目标运动所产生的干扰使磁场发生变化，引起磁强针偏转和摆动，而后将磁信号转换成电信号，从而实现对目标的探测。

磁性传感器鉴别目标性质的能力较强，能准确地探测运动的铁磁金属物体或携带铁磁金属的目标。同时，对目标探测的反应速度也比较快，比前述的两种传感器快得多，可实时地探测快速运动的目标。磁性传感器能适应各种条件下的战场探测，特别是适用于振动传感器难以探测的沼泽、滩涂、水网等地区，从而弥补了振动传感器的不足。由于其受电源和体积的限制，所建立的磁场范围、磁场强度、探测和监视的范围都较小。一般对武装人员的探测距离为5 m以内，对轮式车辆的探测距离为15 m以内，对履带车辆的探测距离为25 m以内。

4. 红外传感器

红外传感器能够感应目标辐射的红外线，并将其转换成电信号。这种传感器通常分为有源式和无源式两种。有源式红外传感器的工作原理是：当战场上运动的人员或车辆通过传感器的工作区域时，传感器发出的红外线即被切断，传感器便被启动，同时监控站的警报器便自动报警。无源式红外传感器的工作原理是：当目标发出热辐射使传感器工作区域的温度突然发生变化时，传感器便被启动。这种装置非常灵敏，在15 m的范围内，人的正常体温的热辐射就足以启动该装置。红外传感器通常隐蔽地布设在需要监视的道路和目标区附近，可探测到视角扇面区20 m以内的人员和50 m以内的车辆目标。

红外传感器的优点是：体积小；隐蔽性好；反应速度快；能探测快速运动的目标；能测定目标的具体方位。其缺点是只能进行人工布设；探测范围只局限于探测器正面的扇形区内；无目标分辨能力；使用范围受到较大的限制。

5. 应变电缆传感器

应变电缆传感器的探测器为一根极细的应变金属丝，由镍铬合金、铁铬合金或康铜（即含40%镍和1.5%锰的铜合金）等金属材料拉制而成，封装入应变电缆。当运动目标通过浅埋在地下的应变电缆时，电缆因受挤压，使其中的应变金属丝变形（伸长或缩短），引起电阻值变化，从而产生一个电信号。

应变电缆传感器只有在运动目标直接碾压应变电缆时才能进行探测，故其探测距离也很小，通常为30 m左右，也就是应变电缆的长度，而且只能人工埋设，野战使用时受限制较多。但是，这种传感器在边防、海防、公安及一些特殊设施的预警工作中使用起来却很方便，效果也很好，其传感响应速度很快。

另外，目前正在研制的地面传感器还有如下几种。

（1）智能传感器。即带有微处理机，兼有判断与信息处理能力的传感器，可对探测值进行修正和误差补偿，提高探测的准确度和可靠性；

（2）光导纤维传感器。即利用光纤本身的某种敏感特性或功能，探测战场指定区域的压力、电场、磁场等环境的变化，以判断敌方目标的性质和活动情况；

（3）图像传感器。即利用光电器件的光电转换功能，将所成的像转换为相应的电信号图像，用以观察战场上声像并存的敌方活动情况；

（4）微量气体传感器。即通过敌方车辆排出气体的气味和含量浓度，判断其车辆的种类和数量。

运用地面传感器进行战场侦察，通常利用一定数量的各类传感器和监视器组成传感器系统。当要进行远距离战场侦察与监视时，还需要在中间加设地面或空中中继器，负责转发信号和指令。传感器串由三个或三个以上的传感器组成，布设在敌人可能活动的区域。传感器区由两个或两个以上的传感器串组成，用来完成特定的任务。

要发挥传感器的优越性，就需要将不同类型和不同发射频率的传感器混合使用。例如，振动传感器和磁性传感器一起使用，就是一种较好的混合使用方法。在这种情况下，振动传感器探测到地表面的振动后，再由磁性传感器探测到该区域内的铁磁金属物体的运动，可起到进一步证明目标性质的作用。一种好的混合式传感器系统能探测和确定入侵的车辆或人员，并能确定车辆或人员的大致数量、纵队的长度、行进方向和运动速度等。

2.6　卫星导航定位系统与技术

卫星导航定位系统由若干颗工作卫星和备用卫星、地面控制系统及用户定位设备等组成。卫星分布在几个不同轨道上，并装备有导航电文存储器、伪码发生器、发射机和接收机等设备。用微波播发的导航电文经接收机处理后，可使全球任何地点和近空间的用户得到全天候、高精度、连续实时的三维坐标，可为战场各作战单元提供用于确定自己所处位置的精确坐标，并为电子系统提供精确的时间标准。

卫星导航定位系统是有史以来最精确的无线电导航系统，在军事上已从单纯的导航定位扩展到目标捕获、武器校射、武器制导、传感器布设、照相侦察、电子情报标注、指挥控制、搜索与救援等各个领域，应用日益广泛。当前世界有四大全球卫星导航系统（Global Navigation Satellite System，GNSS）：美国的全球卫星导航定位系统（Global Positioning System，GPS），俄罗斯的格洛纳斯卫星导航定位系统（Global Navigation Satellite System，GLONASS），欧洲的伽利略卫星导航系统（Galileo）和我国的北斗卫星导航系统。

2.6.1　卫星导航定位系统的基本组成

任何一个卫星导航定位系统至少由三大部分组成，分别是：空间星座部分、地面监控部分和用户设备部分（见图2.10）。

1. 空间星座部分

空间星座由分布在若干轨道面内的若干卫星组成，其信号可以覆盖全球或一定区域。通常星座设计要满足用户

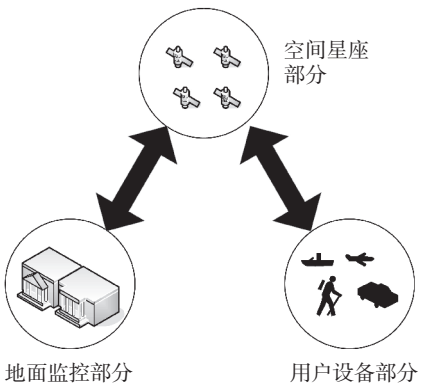


图2.10　卫星导航定位系统组成图

需要导航服务时能够观测到实现卫星导航功能所需的最少卫星数。例如，伪距导航需同时观测4颗卫星。

导航卫星的基本功能是：

(1) 接收和存储由地面监控站发来的导航信号，接收并执行监控站的控制指令；

(2) 卫星上设有计算机，进行必要的数据处理工作；

(3) 通过星载的高精度铯钟和铷钟提供精密的时间标准；

(4) 向用户发送导航信息；

(5) 在地面监控站的指令控制下，通过推进器调整卫星的姿态，或启用备用卫星。

2. 地面监控部分

地面监控部分按照功能可分为三类：监测站、主控站和注入站。

监测站是在主控站直接控制下的数据自动采集中心。站内设有双频测量型接收机、高精度原子钟、计算机和若干环境数据传感器。接收机对导航卫星进行连续观测，以采集数据和监测卫星的工作状态。原子钟提供时间标准，环境数据传感器收集当地的气象数据。所有观测数据由计算机进行初步处理，并传送到主控站进行存储和综合处理，以确定卫星的轨道。

主控站除了协调和管理所有地面监控系统的工作以外，其主要任务如下。

(1) 根据本站和其他监测站的所有观测资料，编制各卫星的星历，推算卫星钟差和大气层修正参数，并把这些数据传送到注入站。

(2) 提供整个系统的时间基准。各监测站和卫星的原子钟均应与主控站原子钟同步，或测出其钟差，并把钟差信息编入导航电文送到注入站。

(3) 调整偏离轨道的卫星，使之沿预定的轨道运行。

(4) 启用备用卫星以代替失效的工作卫星。

注入站的主要设备一般包括C波段发射机和计算机。其主要任务是在主控站的控制下，将主控站推算和编制的卫星星历、钟差、导航电文和其他控制指令等，注入相应卫星的存储系统，并监测注入信息的正确性。

3. 用户设备部分

卫星导航定位系统的空间星座部分和地面监控部分是用户实现导航定位的基础。然而，用户还必须通过用户设备才能实现导航定位的目的。用户设备指的是卫星导航接收机，其主要任务是：接收导航卫星发射的无线电导航信号，以获得必要的定位信息及观测量，并经数据处理完成导航定位工作。

2.6.2 卫星导航信号

卫星播发的导航信号是卫星导航接收机进行导航定位的基础。根据不同导航原理设计的卫星导航系统，其所能提供的导航信号也不尽相同。例如，美国的子午卫星导航系统和我国的北斗一号卫星导航系统的导航信号主要由载波和调制在载波上的导航电文组成；美国的GPS、俄罗斯的GLONASS等卫星导航系统的导航信号由载波及调制在载波上的导航电文、测距码等组成。

1. 载波

根据物理学中的概念，电磁波是一种随时间变化的正弦（或余弦）波。实际上，电磁波传播的空间并不是真空，而是充满大气冷介质的空间。导航卫星发射的电磁波信号到达地面用户

设备之前，要穿过性质和状态各异且不稳定的若干大气层，这可能会改变电磁波传播的方向、速度和强度，这种现象称为大气折射。

根据大气物理学中的概念，电磁波在大气中传播时，大气介质的内电场和入射电磁波的外电场之间会产生电磁转换效应。当大气介质的原子频率与入射电磁波的频率接近一致时，会发生谐振吸收，继而影响电磁波的传播。

地球大气中的氧气和水蒸气对电磁波影响较大。当电磁波工作波长为0.25 cm和0.5 cm时，氧气对其产生谐振吸收而发生最大衰减。当工作波长为0.18 cm和1.25 cm时，水蒸气对其产生谐振吸收而发生最大衰减。

在卫星导航定位中，载波信号一般按其频率（或波长）划分为不同的波段（见表2.3）。现代的卫星导航系统的载波频率一般选择L波段。因为这样可避开大气层中氧、水蒸气等的最大谐振吸收，有利于较经济地接收卫星导航信号，降低卫星信号发播功率，从而减小卫星的功耗。另外，从地面用户设备的角度讲，可降低用户设备功耗和信号接收灵敏度的要求。

表2.3　电磁波波段划分

符号	频率（GHz）	平均波长（cm）
P波段	220~300	115
L波段	1~2	20
S波段	2~4	10
C波段	4~8	5
X波段	8~12.5	3
Ku波段	12.5~18	2
K波段	18~26.5	1.35
Ka波段	26.5~40	1

在卫星导航定位系统中，经常采用扩频通信技术。“扩频”是指将原拟发送的几十比特速率的电文变换成发送几兆甚至几十兆比特速率的电文，再由电文和伪随机噪声码组成组合码。

根据信息论的香农定理，在高斯白噪声干扰条件下，通信系统容量为

$$C = B\log_2\left(1+\frac{S}{N}\right) \tag{2.9}$$

其中，B为通信系统的频带宽度，S为信号的平均功率，N为噪声功率。

由此可见，当系统容量C一定时，增大频率带宽B即可减小信噪比。例如，在上述情况下，C = 10.23 Mbit/s，当信号功率S为噪声功率N的1.5倍时（常用$S > N$甚至$S >> N$），系统的带宽为

$$B = \frac{C}{\log_2(1+S/N)} = \frac{10.23\times10^6}{\log_2(1+1.5)} = 7.74\ \text{MHz} \tag{2.10}$$

如果信号功率仅为噪声功率的十分之一，即信号深深地淹没在噪声中，此时按上式计算可得带宽B = 74.40 MHz。由此可见，可以用增大系统带宽的方法，降低所要求的信噪比，或者说用很小的发射功率便可实现遥远的卫星导航定位。这对于电能紧张的导航卫星极为有益。而且，信号深埋在噪声中，不易被他人捕获，从而具有极强的保密性。因此，卫星导航定位系统采用扩频技术的目的在于：节省卫星的电能；增强卫星导航定位信号抗干扰性；实现保密的信息传递。

2. 导航电文

导航电文就是包含有关卫星的星历、卫星工作状态、卫星时间信息、卫星钟运行状态、卫星轨道摄动修正和实现导航定位所必需的其他信息，是利用卫星进行导航的数据基础。卫

星导航系统把卫星位置作为已知数据，所以导航电文中还包括卫星的位置。导航电文是一种二进制码，按导航系统设计的格式调制在载波上，按帧向外播发。

3. 测距码

测距码是一种调制在载波信号上，用于测量卫星至地面用户设备之间距离的二进制码。由于实现导航定位的原理不同，所以并非所有卫星导航系统播发的导航信号中都有测距码。例如，美国的子午卫星导航系统和我国的“北斗一号”卫星导航系统，所播发的导航信号中都没有在载波上调制测距码。但是，目前应用最为广泛和成功的GPS卫星导航系统、GLONASS系统、正在建设的伽利略系统以及我国“北斗二号”等的导航信号中都有测距码。

2.6.3 伪距卫星导航定位原理

卫星导航定位的实现，必须获取观测站与卫星之间的距离，通过求解观测方程来获取用户当前时刻的状态。由于信号从遥远的卫星经过大气层到达接收机，所测距离难免会受各种误差的影响，所以接收机测出的并非真实距离，而是包含了误差的伪距。

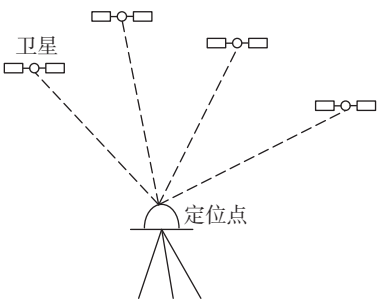


图2.11 卫星定位原理示意图

如图2.11所示，假设卫星S^i播发信号的时刻为T^j，接收机接收信号的时刻为T_R，T表示系统时间，则信号传播时延为

$$\Delta T = T_R - T^j \tag{2.11}$$

卫星至接收机的距离为

$$\rho^j = c \cdot \Delta T = c \cdot \left(T_R - T^j\right) \tag{2.12}$$

实际上，我们只能得到该信号播发时刻卫星钟的钟面时t^j和接收时接收机的钟面时t_R，则所得到的观测量为$\Delta t = t_R - t^j$。

在卫星导航中，时钟钟面时与系统时之差称为钟差。那么，在卫星播发信号时刻T^j，卫星钟差为$\delta t^j = t^j - T^j$；在接收机接收信号时刻T_R，接收机钟差为$\delta t_R = t_R - T_R$，那么，可得如下关系式：

$$t^j = T^j + \delta t^j,\ t_R = T_R + \delta t_R \tag{2.13}$$

设该接收机接收信号时刻与卫星播发信号时刻之差为τ'，则

$$\tau' = t_R - t^j \tag{2.14}$$

将式(2.13)代入式(2.14)，可得

$$\tau' = T_R + \delta t_R - (T^j + \delta t^j) \tag{2.15}$$

$$\tau' = (T_R - T^j) + \delta t_R - \delta t^j \tag{2.16}$$

两边同时乘以光速c，可得

$$c \cdot \tau' = c \cdot (T_R - T^j) + c \cdot \delta t_R - c \cdot \delta t^j \tag{2.17}$$

将式(2.12)代入式(2.17)，并令$\rho'^j = c \cdot \tau'$，经整理可得

$$\rho'^j = \rho^j + c \cdot (\delta t_R - \delta t^j) \tag{2.18}$$

通常，将ρ'^j称为伪距。

设卫星播发信号时刻的卫星坐标为(X^j, Y^j, Z^j)，接收机接收信号时刻的天线中心坐标为(X_R, Y_R, Z_R)，则卫星与接收机之间的几何距离为

$$\rho^j = \sqrt{\left(X^j - X_R\right)^2 + \left(Y^j - Y_R\right)^2 + \left(Z^j - Z_R\right)^2} \tag{2.19}$$

代入式(2.18)，可得

$$\rho'^j = \sqrt{\left(X^j - X_R\right)^2 + \left(Y^j - Y_R\right)^2 + \left(Z^j - Z_R\right)^2} + c \cdot \left(\delta t_R - \delta t^j\right) \tag{2.20}$$

式(2.20)中，(X^j, Y^j, Z^j)和δt^j可通过卫星导航电文提供的卫星轨道参数、钟差参数计算得到，ρ^j为观测量，$(X_R, Y_R, Z_R, \delta t_R)$为未知参数。

如果接收机同时观测了4颗卫星，即可求解得到$(X_R, Y_R, Z_R, \delta t_R)$这4个参数。如果$\delta t_R$已知，仅需同时观测3颗卫星即可解得$(X, Y, Z)$，但事实上用户很难以足够的精度测定接收机的钟差，所以一般把它作为一个位置待定参数与接收机一并解出。所以，需要至少同时观测4颗卫星，即

$$\rho'^j = \sqrt{\left(X^j - X_R\right)^2 + \left(Y^j - Y_R\right)^2 + \left(Z^j - Z_R\right)^2} + c \cdot \left(\delta t_R - \delta t^j\right), \quad j = 1, 2, 3, 4 \tag{2.21}$$

以上即为伪距卫星导航定位的基本原理。伪距卫星导航定位是现代卫星导航定位中最常用的定位方式，在GPS、GLONASS和伽利略等卫星导航系统中得到了广泛应用。

2.6.4　伪随机测距码

目前卫星导航系统中采用的测距码都是伪随机码。20世纪40年代末期，香农首先指出白噪声形式的信号是一种实现有效通信的最佳信号，但因产生、加工、控制和复制白噪声存在困难，香农的设想未能实现。直到20世纪60年代中期，随着伪随机噪声码编码技术的问世，噪声通信才获得了实际的应用，并随即扩展到了雷达和导航等技术领域。

1. 码的概念

码是指一种表达信息的二进制数及其组合，是一组二进制的数码序列。如果将各种信息按某种预定的规则表示为二进制数的组合，则称这一过程为编码。例如，若将地面测量控制网分为4级，并用二进制数表示，则可取两位二进制数的不同组合，即使用11、10、01和00依次代表控制网的1、2、3、4级。这些二进制数的组合形式就是码。其中每个码均含有2个二进制数，即2个比特（Binary digit，简称Bit）。比特的单位也是码的度量单位。

在二进制的数字化信息传输中，每秒所传输的比特数就是码速率，用于表示数字化信息的传输速度，其单位为bit/s，或记为bps。

(1) 码元

对应矩形波出现一次+1或−1，称为一个码元。

(2) 码元长

一个码元对应的时间长，也就是传送一位二进制码所需的时间，通常可记为纳秒。

2. 伪随机噪声码

伪随机噪声码（Pseudo Random Noise Code，PRN码）是一种可预先确定并能够重复产生和复制，具有随机统计特性的二进制码序列，也称为伪码、伪随机码或伪随机噪声码。

伪随机噪声通信采用伪随机噪声码，卫星导航所用的伪随机噪声码是伪随机噪声通信的成功实践。所谓伪随机噪声码，简而言之，是一个具有周期性的取值为0或1的离散符号串，具有

类似于白噪声的自相关函数。伪随机噪声码可以实现低信噪比接收，并且可实现码分多址通信，此外伪随机噪声码还有很高的保密性。

目前的卫星导航系统一般采用一种易于产生、应用广泛的伪随机噪声码——最长线性移位寄存器序列，简称*M*序列。*M*序列是由若干级带有特定反馈电路的移位寄存器产生的。

3. 测码伪距

无线电测距的基本原理是，通过测定电磁波传播时间 τ 得到距离观测量

$$\rho = c \cdot \tau$$

其中，*c*为电磁波传播速度。

卫星导航系统在利用伪随机噪声码测距时，均基于这个原理。接收机利用本机产生的与发射信号相同的复现信号（称为本地信号），与所接收到的含有噪声的信号进行相关处理，并测量相关函数最大值的位置，使得本地码与接收信号中的码同步，通过监测本地码相位的变化实现距离测量。

（1）伪随机噪声码测距的基本思想

如图2.12所示，从接收机在*t*时刻接收到的卫星信号中任取一个长度为码周期的信号，该信号中包含伪随机噪声码$C(t)$，其中码相位为 τ 。对GPS的C/A码来说，周期$T_{C/A}$为1 ms，共包含1023个码元，相位 $\tau = N/1023 (N = 0, 1, \cdots, 1023)$。假设卫星发送信号时刻该伪随机噪声码的相位为0，信号经传播时延 $\tau + nT_{C/A}$ 后到达接收机。显然，发送时刻信号中包含的伪随机噪声码为$C(t - \tau - nT_{C/A})$，*n*为正整数。

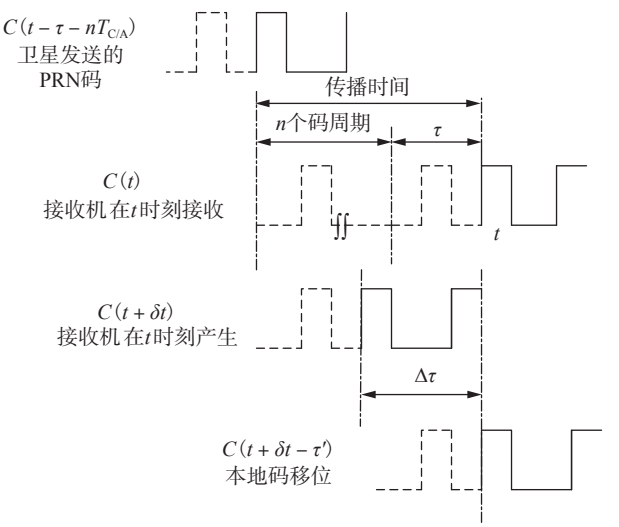


图2.12 伪随机噪声码相位测量

由接收机时钟控制的本地码发射器产生一个与接收到的伪随机噪声码相同类型的本地码$C(t + \delta t)$，初始相位为0，δt 为接收机时钟相对于卫星时钟的钟差。将本地码移位（延迟）τ'，得到$C(t + \delta t - \tau')$。

将接收码$C(t - \tau - nT_{C/A})$和本地码$C(t + \delta t - \tau')$进行相关运算，经积分器后可得相关输出，即

$$R_c(\Delta\tau) = \int C\left(t - \tau - nT_{C/A}\right) C\left(t + \delta t - \tau'\right) \mathrm{d}t \tag{2.22}$$

其中

$$\Delta\tau = (t + \delta t - \tau') - (t - \tau - nT_{C/A}) \tag{2.23}$$

码振荡器不断调整本地码延迟，当本地码与接收码完全对齐时 $\Delta\tau = 0$，且相关输出达到最大值

$$\left[R_c(\Delta\tau)\right]_{\max} = R_C(0) \tag{2.24}$$

假设此时本地码的相位延迟量为 τ'，则由式(2.23)可得

$$t+\delta t-\tau'=t-\tau-nT_{\mathrm{C/A}}\Rightarrow\tau'=\tau+\delta t+nT_{\mathrm{C/A}} \tag{2.25}$$

式(2.25)两边同时乘以电磁波传播速度c，可得距离测量值ρ为

$$\rho=c\cdot\tau'=c\cdot(\tau+\delta t+nT_{\mathrm{C/A}})=R+c\cdot\delta t \tag{2.26}$$

上式为伪随机噪声码测距的基本方程。其中，R为卫星至接收机的真实距离。

上述推导过程中将卫星时钟视为完全精确。实际上，相对于导航系统基准时，卫星时钟同样存在钟差δt_s，如果假设接收机时钟相对于导航系统基准时的钟差为δt_k，则式(2.26)中，$\delta t=\delta t_k-\delta t_s$。因此，利用伪随机噪声码所测定的距离$\rho$并不等于信号从卫星传播至接收机的真实距离，若能精确求得δt，则可由ρ求得R。实际上，δt_s可由导航电文提供的卫星时钟误差参数计算得出，而接收机通常为节省成本采用了廉价的时钟，无法达到原子钟的精度，误差较大，只能在定位解算中将其作为一个待定参数求解。因此，利用测码伪距定位，需要至少同步观测4颗卫星。测相伪距中同样会存在这个问题。

(2) 获取码伪距

由于在式(2.25)中，将δt作为未知数处理，也就是说，信号接收时间为本地时间，那么，接收机只需测得精确的信号发送时间即可完成伪距测量。

GPS的每一子帧电文持续时间为6 s，共有300位，每位电文持续20 ms，对应20个C/A码周期。在卫星上，子帧中的Z计数值从周日零时以后以6 s间隔从1开始计数，同时卫星开始发送第一子帧电文。因此，任一子帧的Z计数值减1后乘以6 s即表示当前子帧起始位置的发送时间。

如图2.13所示，当接收机在某个本地时刻需要测量伪距时，首先从解调的导航电文中获得Z计数，得到当前子帧起始位置对应的发送时间。如果能够获得Δt，即可得出伪距测量时刻电文的发送时间。

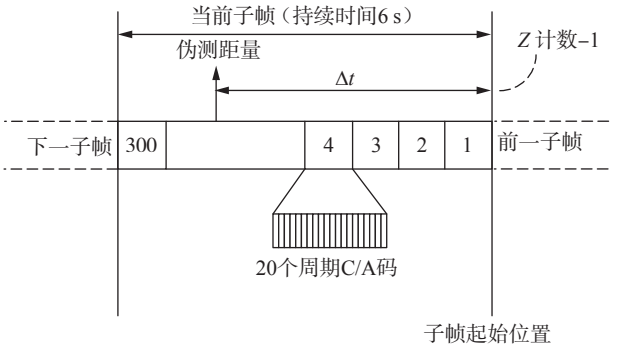


图2.13　获取码伪距

由于Δt由周期为20 ms的导航电文(N_{Data})、整周期的C/A码($N_{\mathrm{C/A}}$)、不足1周期的C/A码片数(N_{Chips})以及不足1个码片的部分(N_{Rcs}, $0<N_{\mathrm{Rcs}}<1$)构成：

$$\Delta t=N_{\mathrm{Data}}\times20\times10^{-3}+N_{\mathrm{C/A}}\times10^{-3}+\frac{N_{\mathrm{Chips}}}{1.023\times10^{6}}+\frac{N_{\mathrm{Rcs}}}{1.023\times10^{6}} \tag{2.27}$$

因此，如果接收机能够对接收信号中的导航电文和伪随机码进行计数，就可以获得精确的信号发送时间。由于接收机能够利用码跟踪环使本地码与接收码精确同步，所以对本地码的计数间接地实现了对接收信号中的伪随机码的计数，再综合码振荡器的变化和解调的电文位数即可获得Δt。Δt加上Z计数的时间，即为信号的精确发送时间，再与本地接收时间求差即可得到信号的传播时间。

采用伪随机码测距技术时，距离测量误差在很大程度上取决于本地码与接收机码的对准误差，而码相位对准误差主要受热噪声、载体动态性和多径干扰的影响。由于接收机采用诸如窄相关时间、码相位拟合等技术，相位对准误差远小于1个码片，对于GPS的C/A码，测距误差通常在0.3~30 m的范围内。

2.6.5 四大全球导航定位系统

1. 美国全球卫星导航定位系统（GPS）

为了满足军事部门和各种民用部门对连续实时定位和三维导航的迫切要求，1973年美国国防部便开始组织海陆空三军，共同研究建立新一代卫星导航系统的计划。这就是目前所称的“导航卫星测时测距/全球定位系统”，而通常简称为“全球定位系统（GPS）”。

GPS主要由三大部分构成：空间部分——GPS卫星星座；地面控制部分——地面监控系统；用户设备部分——GPS信号接收机。

（1）GPS卫星星座

GPS卫星星座配置如图2.14所示，由24颗卫星组成，包括21颗工作卫星和3颗在轨备用卫星，记为(21 + 3) GPS星座。卫星均匀分布在6个轨道平面内，每个轨道面上分布了4颗卫星。卫星轨道面相对于地球赤道面的倾角约为55°，各个轨道平面之间相距60°，即轨道的升交点赤经各相差60°。每个轨道平面内各颗卫星之间的升交角距相差90°，一轨道平面上的卫星比西边相邻轨道平面上的相应卫星超前30°。轨道平均高度约为20 372 km，卫星运行周期为11小时58分钟。

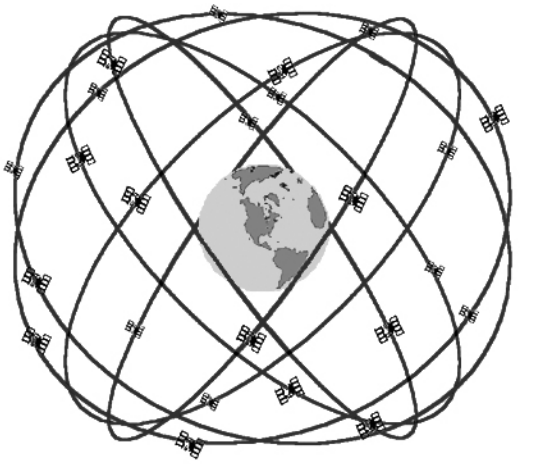


图2.14 GPS星座配置图

对于两万千米高空的GPS卫星，当地球对恒星来说自转一周时，它们绕地球运行两周，即绕地球一周的时间为12恒星时。这样，对于地面观测者来说，每天将提前4分钟见到同一颗GPS卫星。位于地平线以上的卫星颗数随着时间和地点的不同而不同，最少可见到4颗，最多可见到11颗。在用GPS信号导航定位时，为了解算测站的三维坐标，必须观测4颗GPS卫星，称为定位星座。这4颗卫星在观测过程中的几何位置分布对定位精度有一定的影响。对于某地某时，甚至不能测得精确的点位坐标，这种时间段称为“间隙段”。但这种时间间隙段是很短暂的，并不影响全球绝大多数地方的全天候、高精度、连续实时的导航定位测量。

（2）地面监控系统

对于导航定位来说，GPS卫星是一动态已知点。星的位置是依据卫星发射的星历——描述卫星运动及其轨道的参数算得的。每颗GPS卫星所播发的星历，是由地面监控系统提供的。卫星上的各种设备是否正常工作，以及卫星是否一直沿着预定轨道运行，都要由地面设备进行监测和控制。地面监控系统的另一重要作用是保持各颗卫星处于同一时间标准——GPS时间系统。这就需要地面站监测各颗卫星的时间，求出钟差，然后由地面注入站发给卫星，卫星再由导航电文发给用户设备。GPS工作卫星的地面监控系统包括一个主控站、三个注入站和五个监测站。

（3）GPS信号接收机

GPS信号接收机的任务是：能够捕获到按一定卫星高度截止角所选择的待测卫星的信号，并跟踪这些卫星的运行，对所接收到的GPS信号进行变换、放大和处理，以便测量出GPS信号从卫星到接收机天线的传播时间，解译出GPS卫星所发送的导航电文，实时地计算出测站的三维位置，甚至三维速度和时间。

GPS卫星发送的导航定位信号，是一种可供无数用户共享的信息资源。对于陆地、海洋和空间的广大用户，只要用户拥有能够接收、跟踪、变换和测量GPS信号的接收设备，即GPS信号接收机，就可以在任何时候用GPS信号进行导航定位测量。根据使用目的不同，用户要求的GPS信号接收机也各有差异。目前世界上已有几十家工厂生产GPS接收机，产品也有几百种。这些产品可以按照原理、用途、功能等来分类。

静态定位时，GPS接收机在捕获和跟踪GPS卫星的过程中固定不变，接收机高精度地测量GPS信号的传播时间，利用GPS卫星在轨的已知位置，解算出接收机天线所在位置的三维坐标。而动态定位则是用GPS接收机测定一个运动物体的运行轨迹。GPS信号接收机所位于的运动物体称为载体（如航行中的船舰、空中的飞机、行走的车辆等）。载体上的GPS接收机天线在跟踪GPS卫星的过程中相对地球而运动，接收机用GPS信号实时地测得运动载体的状态参数（瞬间三维位置和三维速度）。

接收机硬件和机内软件以及GPS数据的后处理软件包，构成完整的GPS用户设备。GPS接收机的结构分为天线单元和接收单元两大部分。对于测地型接收机来说，两个单元一般分成两个独立的部件，观测时将天线单元安置在测站上，接收单元置于测站附近的适当地方，用电缆线将两者连接成一个整机。也有的将天线单元和接收单元制作成一个整体，观测时将其安置在测站点上。目前，各种类型的GPS接收机体积越来越小，重量越来越轻，便于野外观测。

相对于其他导航系统，GPS具有以下几个特点。

（1）全球地面连续覆盖。由于GPS卫星的数目较多，且分布合理，所以地球上任一地点、任一时刻均可观测到至少4颗以上的卫星，从而保证了全球、全天候的三维连续定位。

（2）功能多，精度高，操作简便。GPS可为各类用户连续地提供动态目标的三维位置、三维速度和时间信息。在工程测量中，300~1500 m的工程精密定位中，对于观测1 h以上的结果，其平面误差小于1 mm，有时甚至可小于0.5 mm。

（3）实时定位，全天候作业。利用GPS导航，可以实时地确定运动目标的三维位置和速度，由此既可以保障运动载体沿预定航线飞行，也可实时监视和修正航行路线，以及选择最佳的航线。

随着GPS定位技术的发展，其应用的领域在不断扩宽。目前，GPS系统的应用已经十分广泛，我们可以应用GPS信号进行海、空和陆地的导航，导弹的制导，大地测量和工程测量的精密定位，时间的传递和速度的测量等。对于测绘领域，GPS卫星导航定位技术已经用于建立高精度的全国性的大地测量控制网，测定全球性的地球动态参数；用于建立陆地、海洋、大地测量基准，进行高精度的海岛、陆地联测以及海洋测绘；用于监测地球板块运动状态和地壳形变；用于工程测量，成为建立城市与工程控制网的主要手段；用于测定航空航天摄影瞬间的相机位置，实现仅有少量地面控制或无地面控制的航测快速成图，从而实现了地理信息系统、全球环境遥感监测的技术革命。

2. 俄罗斯卫星导航定位系统（GLONASS）

在美国发展GPS的同一时期，前苏联也独立研制了性能与GPS基本相同的卫星导航系统GLONASS。并在苏联解体后由俄罗斯继续完成了星座布置，保证系统完全布满后至少在15年内提供全球服务。该系统的24颗卫星定轨于3个圆轨道平面上，每个平面包括8颗卫星，运行周期11小时15分钟，离地高度19 100 km，轨道倾角64.8°，总体布局与GPS星座相当。

21世纪初俄罗斯因一度没有及时为失效的卫星发射替代卫星，使系统的在轨卫星数量急剧下降，严重影响了其使用效能。近年来随着经济状况的好转和卫星导航定位市场的蓬勃发展，俄罗斯已经开始重建GLONASS系统。到2010年俄罗斯使卫星数量达到“满员”状态。

GLONASS系统除采用频分多址（Frequency Division Multiplex Access，FDMA）技术与GPS略有不同外，在系统配置、定位机理、工作频段、信号和星历数据结构等方面与GPS系统基本相同，都以发射扩频测距码测量伪距来完成导航定位。GLONASS系统性能与GPS基本相同，它的主要特点是该系统能为飞机、航海和其他用户提供坐标定位和测速等导航信息，并能提供全球覆盖；该系统的导航精度较高；可以为所有用户提供服务，无任何限制，对最终用户没有收费计划，因此具有良好的应用前景。

GLONASS的建成和公开化，打破了美国对卫星导航独家经营的局面，既可为民用用户提供独立的导航服务，又可与GPS组合提供更好的导航性能，同时它也大大缓解了美国政府利用GPS施以主权威慑给用户带来的后顾之忧。GLONASS与GPS为全球卫星导航系统在世界范围内得到广泛应用开辟了美好的前景。

3. 欧洲伽利略卫星导航系统（Galileo）

2005年12月28日，欧洲伽利略卫星导航系统（Galileo）的首颗实验卫星“GIOVE-A”顺利进入太空，这是欧洲为打破美国在卫星导航系统上的垄断局面所迈出的重要一步，标志着伽利略全球卫星导航系统进入了正式的轨道验证阶段。

伽利略系统是欧洲计划建设的新一代民用全球卫星导航系统，系统星座计划由30颗卫星组成，其中包括27颗工作卫星和3颗备用卫星。卫星采用中等地球轨道，均匀分布在高度约为23 000 km的3个轨道面上，每个轨道10颗卫星，其中1颗为当前轨道的备用卫星。

由欧盟和欧洲空间局共同提出和组织实施的伽利略计划，旨在建立一个民用的全球卫星导航系统。伽利略系统在整个建设和维护过程中没有军方直接参与，这是伽利略系统首先区别于GPS和GLONASS系统的特点。另外，伽利略系统在总体构成、频段划分、信号设计和安全保障措施等方面均与其他卫星导航系统有所不同，其特点主要有以下几个方面。

（1）伽利略系统从概要设计开始就考虑了系统的完好性问题，从系统结构设计的角度对完好性加以处理，保证了全球的可用性。

（2）伽利略星座的卫星数量多，轨道位置高，轨道面少，覆盖面积大。即使是处于地球两极高纬度地区的用户，也能获得较好的定位结果。

（3）伽利略系统能够提供多样化、多层次的服务：开放服务、生命安全服务、商业服务、公共特许服务和搜救服务，能够满足全球用户的需要，真正体现民用卫星导航系统的价值。

（4）伽利略系统既与现有的卫星导航系统保持独立又具有兼容性，欧盟已与美国达成两个系统互相兼容的协议，两者信号在混合使用的情况下，定位精度将更高，同时也为多频段、多星座的卫星导航接收机提供了发展空间。

（5）伽利略系统可为移动电话业务提供服务，能用于搜索救援，为对生命安全要求高的应用提供服务。

（6）伽利略系统允许其他非欧盟国家为系统的建设提供资金和技术支持，共享伽利略系统提供的各项服务。

4. 北斗卫星导航系统

北斗卫星导航系统由我国自主建立，以“先区域，后全球”的建设思想分为“北斗一号”和“北斗二号”两个阶段。“北斗一号”卫星导航系统是具备通信功能的、区域性有源定位双星

导航系统，能实现中国和东南亚地区的导航、通信、授时服务。“北斗一号”于2003年正式投入使用以来，工作状态稳定可靠，并逐步向“北斗二号”全球卫星导航系统过渡。

（1）“北斗一号”卫星导航系统

2003年5月25日，我国成功发射了第三颗“北斗一号”导航定位卫星，作为北斗导航定位系统的备份星，连同2000年10月31日和12月21日发射升空的两颗“北斗一号”导航定位卫星和一个地面中心站，形成了一个较为完善的双星导航定位系统。它是我国第一代卫星导航定位系统，是我国独立自主开发的可以提供全天候、高精度、大范围、快速实时定位的区域性卫星导航定位系统。由此我国成为世界上继美国、俄罗斯之后，第三个拥有卫星导航系统的国家。

“北斗一号”卫星导航定位系统的结构与GPS系统类似，也是由卫星星座、地面控制站和用户接收机三部分组成的。与全球卫星导航系统不同的是，“北斗一号”只有三颗星，其中两颗工作星，一颗备用星，属于区域卫星导航系统。

（2）定位原理

“北斗一号”导航定位系统由两颗经度上相距60° 的静止卫星，一个配有电子高程图的地面中心站，几十个分布于全国的参考标校站和大量用户机组成。它的定位原理是：以两颗卫星的已知坐标为圆心，各以测定的本星至用户机距离为半径R，形成两个球面，用户机必然位于这两个球面交线的圆弧上。电子高程地图提供的是一个以地心为球心，以球心至地球表面各个目标高度为半径的非均匀球面。求解圆弧线与地球表面交点即可获得用户的位置。具体的定位过程是：首先由地面中心发出信号，分别经两颗卫星反射传至用户接收机，再由接收机反射经两颗卫星分别传回地面中心，地面中心站计算出两个途径所需时间t_1和t_2，设卫星的位置为$R_1(x_1, y_1, z_1)$和$R_2(x_2, y_2, z_2)$，R_1和R_2可由地面中心确定，通过下列方程组和电子高程地图就可以计算待测点位置(x, y, z)：

$$ct_1 = 2(\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2} + R_1)$$

$$ct_2 = 2(\sqrt{(x_2 - x)^2 + (y_2 - y)^2 + (z_2 - z)^2} + R_2)$$

其中，c为电磁波的空中传播速度，即光速。

上述一系列复杂的运算，对“北斗一号”局域导航系统来说，是在地面中心站进行的。地面中心站确定用户位置后，再把定位与时钟信息通过卫星传给用户。“北斗一号”导航定位系统也可确定用户的运动速度，因为中心站存储了用户的新旧数据，把新旧数据一对比就能立即求出用户的运动速度。

“北斗一号”的定位精度不低于GPS的C/A码精度。

（3）“北斗二号”卫星导航系统

与采用被动定位方式实现的全球性卫星导航系统相比，采用主动定位方式的“北斗一号”由于卫星数量有限，在信号覆盖范围、定位精度、隐蔽性、系统容量等方面存在很多不足，已不能满足我国日益增长的导航需求，其他卫星导航系统的发展也对“北斗一号”提出了更高的挑战。

为了克服“北斗一号”卫星导航系统的缺点，保留其可以进行报文通信的优点，我国于2004年开始筹建性能更高、覆盖面更广、技术更先进的“北斗二号”全球卫星导航系统。2007年4月和2009年4月先后成功发射两颗“北斗二号”卫星进入预定轨道，标志着系统卫星组网工作正式启动。2012年10月25日，我国成功发射第16颗北斗导航卫星，自此形成覆盖亚太地区的区域卫星导航系统。2012年12月27日，北斗系统空间信号接口控制文件正式版发布，北斗导航

业务正式对亚太地区提供无源定位、导航、授时服务。根据计划，到2020年左右，“北斗二号”将形成全球覆盖能力。

“北斗二号”卫星导航系统的卫星星座计划由5颗地球静止轨道卫星和30颗非静止轨道卫星组成。其中5颗地球静止轨道卫星高度为36 000 km，在赤道上空分布于东经58.75°、东经80°、东经110.5°、东经140° 和东经160°；30颗非静止轨道卫星由27颗中地球轨道卫星和3颗倾斜同步轨道卫星组成；27颗中地球轨道卫星分布在倾角为55° 的3个轨道平面上，轨道高度为21 500 km。

“北斗二号”的地面控制站包括1个主控站、2个注入站和30个监测站。监测站实时跟踪监测卫星工作状况和监测站附近的空间、地理环境的变化，并将这些信息传送给主控站。主控站接收监测站发送的数据，编算导航电文、星历数据，将其与时间基准一同传送至注入站，协调管理注入站和监测站的工作，并根据监测数据控制卫星运行状态，保证“北斗二号”星座正常运转。注入站将卫星星历、导航电文、钟差和其他控制指令注入卫星。

作为北斗第二代卫星导航系统，“北斗二号”既能够兼容“北斗一号”，又与其在工作原理和性能上存在如下明显区别。

(1)“北斗二号”卫星导航系统的接收机可免发上行信号，不再依赖主控站，而是由接收机解算位置坐标，系统的用户容量不受限制，定位隐蔽性提高；

(2) 采用多颗卫星进行定位，而不是双星定位，不需要高程信息辅助；

(3) 保留了“北斗一号”的通信功能，能够实现报文或指令通信；

(4) 定位精度、授时精度更高。

“北斗二号”卫星导航系统建成后将提供两种服务，一种是针对非授权用户的开放服务，另一种是针对特许用户的授权服务。开放服务在信号覆盖区免费提供定位、测速和授时服务，在全球范围内定位精度可达10 m，授时精度可达10 ns，测速精度为0.2 m/s。授权服务可为有高精度、高可靠卫星导航需求的用户提供更安全及更高精度的定位、授时、测速和通信服务，以及系统完好性信息，满足全球范围内用户对导航的需求。局部区域内的差分定位精度可达1 m，并且可以利用“北斗二号”卫星进行报文通信。

思考题

1. 什么是信息获取技术？可以通过哪些手段获取目标信息？
2. 简述无线电波的波段划分和不同波段的主要传播方式。
3. 光电信息获取技术主要包括哪些种类？
4. 红外成像技术与可见光成像技术相比有哪些特点？
5. 简述雷达的组成和工作原理。
6. 简要说明雷达是如何对目标定位和测速的。
7. 相控阵雷达与常规机械扫描雷达相比，工作原理有什么不同？有哪些优点？
8. 声呐按工作方式可分为哪两类？说明其各自的优缺点和应用场合。
9. 什么是伪距？其作用是什么？

第3章　信息传输与交换技术

信息传输与交换技术，是应用信息科学原理和方法，实现并扩展人类传导神经网络器官的功能，消除和克服空间上的限制，使人们能更有效地利用信息资源的技术，也就是我们常说的通信技术。通信系统中涉及的具体技术很多，本章主要介绍基本概念和信息的传输、交换等主要技术。

3.1　通信系统基本概念

如果说，通信主要是指信息的传输与交换过程，那么通信系统则是指完成通信过程的全部设备和传输介质。

3.1.1　通信系统模型

虽然通信系统的形式各式各样，具体设备和业务功能也各不相同，但是经过概括和抽象，可提炼出通用的基本组成模型（见图3.1）。

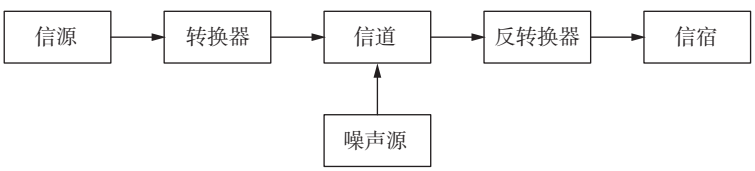


图3.1　通信系统模型

1. 信源

信源是指信息的发送者，可以是发送信息的人或设备（如计算机）。信源所发送的信息按照其表现形式，可分为语音信息、数据信息和图像信息等。不同的信源往往产生不同的通信业务，构成不同的通信系统，如电话通信系统、数据通信系统和图像通信系统等。

2. 转换器

转换器的作用是将信源发送的信息转换成适合信道传输的信号（一般为电信号）。对于不同的信源和不同的通信系统，转换器有不同的组成结构和功能。例如，对于模拟电话通信系统，转换器就是送话器，主要完成声电转换功能；对于数字电话通信系统，转换器则包含了送话器和模数转换器等部件，不仅要完成声电转换功能，而且要进行模数转换、编码以及时分复用等处理，以适用于在数字信道中传输。

3. 信道

信道是信息的传输通道，通常由传输介质和传输设备组成。按照所使用的传输介质的不同，信道可分为有线信道和无线信道。在有线信道中，电磁信号被约束在指定的传输线路中传

输，常用的传输线路有架空明线、被复线、电缆、光缆等。在无线信道中，电磁信号是在开放的空间中传输的。常用的无线信道包括短波无线电信道、超短波无线电信道和微波无线电信道等。在信道中传输的信号大体分为两种形式，一种是模拟信号，另一种是数字信号，相应的信道分别称为模拟信道和数字信道。

4. 反转换器

反转换器的作用与转换器的作用正好相反，它将从信道上接收的信号转换成信宿可以接收的信息形式。例如，对于模拟电话通信系统，反转换器就是受话器，用以完成电声转换功能；对于数字电话通信系统，反转换器则包括受话器和数模转换器等部件，信道中的信号经分解、解码、数模转换和电声转换后还原为可以接收的话音信息。

5. 信宿

信宿即信息的接收者，可以是与信源相对应的人或机器，也可以是与信源不一致的人或机器。根据信宿与信源的对应关系是否一致，可以得到以下四种基本的通信方式：人–人通信、机–机通信、人–机通信和机–人通信。

6. 噪声源

噪声源是指系统中各种干扰和噪声对信息传输的影响在等效意义上的集中体现。通信系统中噪声来源很多，比如通信系统设备中导体本身产生的热噪声、半导体器件的散弹噪声以及来自宇宙的噪声等。总体上讲，系统的各种噪声主要来自内部电子噪声和外部电磁干扰两个方面。这些噪声是分布在系统各个部分的，为了描述和分析问题方便，通常将系统中的内部噪声和外部干扰统一折算到信道中，用一个等效的噪声源来表示，并等效认为它是作用于信道的，这可以大大简化分析。

特别要注意的是，图3.1所示的通信系统，实际上只能完成点到点的单向通信。若要实现双向通信，则还需要增加一个与该图信息传输方向相反的通信系统。若要实现多用户（两个以上用户）之间的通信，则需要具有交换处理功能的通信网，此时由通信网提供信道。

3.1.2 通信系统分类

通信系统有多种分类方法，从通信系统模型的角度来看可以分为以下几种。

1. 按通信业务分类

根据通信业务的类型不同，可以把通信系统分为电话通信系统、电报通信系统、图像通信系统、数据通信系统等。在这些系统中，电话通信系统是最发达的，其他通信系统往往借助于公共的电话通信系统来进行。例如，电报通信通常是从电话话路中划分出一部分频带来传送，或者用一个电话话路传送多路电报。再如，近距离的数据通信可以采用专线形式，但是远距离传输时考虑到成本等原因，往往借助电话信道传送。家庭利用电话线连接国际互联网（Internet）就属于这一种。通信系统的发展趋势是，各种类型的消息能在一个统一的通信网络中传输、交换和处理，即实现通信系统综合化。

2. 按调制方式分类

根据信道中传输的信号是否经过调制，可将通信系统分为基带通信系统和频带通信系统。

基带通信系统是指基带信号（由消息转换成低频电信号）不经调制直接进行传送的通信系统，如音频市内电话、数字基带传输系统等。频带通信系统则是指对各种基带信号调制以后再进行传送的通信系统。调制的原因如下：

(1) 将消息转换为便于传送的形式。如无线传输话音信号必须进行调制，将其加载到高频载波上。

(2) 提高性能，特别是抗干扰能力。比如采用调频技术后，通过提高调频信号的带宽，接收端的信噪比可以有极大的改善。

(3) 有效地利用带宽，实现信道的复用。

常见的调制分类见表3.1。

表3.1　常见的调制方式及用途

调制方式			主要用途
载波调制	线性调制	常规双边带调制（AM）	广播
		单边带调制（SSB）	载波通信、短波无线电话通信
		双边带调制（DSB）	立体声广播
		残留边带调制（VSB）	电视广播、传真
	非线性调制	频率调制（FM）	卫星中继、卫星通信
		相位调制（PM）	中间调制方式
	数字调制	振幅键控（ASK）	数据传输
		频移键控（FSK）	数据传输
		相移键控（PSK）和差分相移键控（DPSK）	数据传输
		其他高效数字调制，如正交振幅调制（QAM）和最小频移键控（MSK）	数字微波、空间通信
脉冲调制	脉冲模拟调制	脉幅调制（PAM）	中间调制方式、遥测
		脉宽调制（PDM）	中间调制方式
		脉位调制（PPM）	遥测、光纤传输
	脉冲数字调制	脉码调制（PCM）	市话中继线、卫星、空间通信
		增量调制（DM）	军用、民用数字电话
		差分脉码调制（DPCM）	电视电话、图像编码

3. 按信号特征分类

如果信道中传输的是模拟信号（时间与取值都连续的信号），则称该通信系统为模拟通信系统；如果信道中传输的是数字信号（时间与取值都离散的信号），则称该通信系统为数字通信系统。

数字通信与模拟通信相比较有很多优点：

(1) 抗干扰能力强。

(2) 差错可控。

(3) 易加密。

(4) 易与计算机、数字处理等现代技术相结合。

4. 按传输介质分类

按传输介质分类，通信系统可分为有线通信系统和无线通信系统两大类。有线通信系统是用导线（如架空明线、同轴电缆、光导纤维及波导等）作为传输介质完成的通信，如市内电话、有线电视及海底光缆通信等。无线通信系统则依靠电磁波在空间传播达到传递消息的目的，如短波电离层传播、微波视距传播、卫星中继等。

5. 按信号复用方式分类

按信号复用方式不同，通信系统分为频分复用（Frequency Division Multiplexing，FDM）通信系统、时分复用（Time Division Multiplexing，TDM）通信系统和码分复用（Code Division Multiplexing，CDM）通信系统。多路信号传输有三种复用方式，即FDM、TDM和CDM。FDM是用频谱搬移的方法使不同信号占据不同的频率范围，TDM是用脉冲调制的方法使不同信号占据不同的时间区间，CDM是用正交的脉冲序列分别携带不同的信号。传统的模拟通信中都采用FDM。随着数字通信的发展，TDM通信系统的应用越来越广泛；CDM主要用于移动通信和空间通信的扩频通信。

3.1.3 通信方式

通信方式是指通信双方之间的工作方式或信号传输方式。信号在信道中传输，可采用如下多种方式。

1. 单工通信、半双工通信、全双工通信

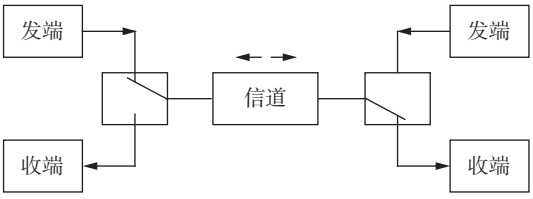
对于点对点通信，按消息传送的方向与时间的关系，通信方式可分为三种：单工、半双工、全双工。

单工通信是消息在任意时刻只能单方向传输的一种通信方式，如图3.2(a)所示。日常生活中单工通信的例子很多，如电视广播、遥控等。这些系统中，信号只能从电视台、遥控器分别向电视机、遥控对象进行单向传送。

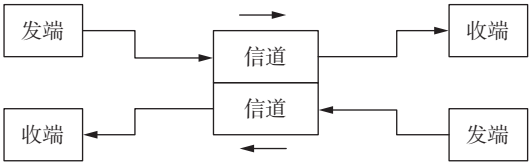
半双工通信是通信双方虽然都能进行收或发信息，但不能同时进行收和发的通信方式，如图3.2(b)所示。例如，对讲机、收发报机等都是这种方式。进行半双工通信时的通信双方，某时刻一方发送时，另一方只能接收。

全双工通信是通信双方可同时进行双向消息传输的一种通信方式，如图3.2(c)所示。全双工通信系统中，通信的双方可同时收发信息，生活中的普通电话、移动电话都是全双工通信方式。

(a)单工通信



(b)半双工通信



(c)全双工通信

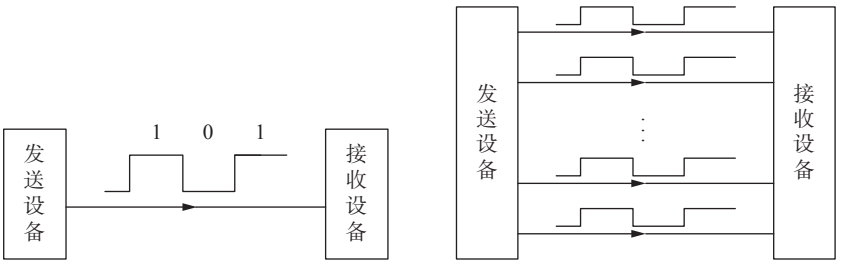
图3.2 通信方式示意图

2. 串行通信与并行通信

在数字通信中，按数字信号码元的排列方法不同，可分为串行通信和并行通信。

所谓串行通信方式，是将代表信息的数字信号或数据信号按时间顺序一个接一个地在信道中传输的方式，如图3.3(a)所示。串行通信方式由于只需要一条通路，所以适用于远距离传输数字信息。

如果将代表信息的数字信号序列按某一规则分成两路或两路以上的数字信号序列，以便同时在信道上传输，则称为并行通信方式，如图3.3(b)所示。并行通信方式与串行通信方式相比具有较高的码元传输速率，但通常只在短距离传输数字信息时使用。



(a)串行通信　　(b)并行通信

图3.3　串行通信和并行通信方式示意图

3.1.4　通信系统的主要性能指标

通信的目的是传递信息，传输信息的有效性和可靠性是衡量通信系统最主要的性能指标。有效性是指通信系统中信息传输的速度大小，而可靠性是指信息传输的质量高低。有效性与可靠性在具体通信系统中有不同的衡量方法。

3.1.4.1　模拟通信系统的性能指标

1. 有效性指标

模拟通信系统的有效性指标用传输带宽来衡量。传输带宽是系统所能传输的信号频率范围。例如，一个2~3 GHz传输带宽的模拟系统，就可以传输2~3 GHz的信号，其他频率的信号不能在该系统中有效传输。

需要说明的是，有三个与传输带宽有关联的概念，包括信号带宽、系统带宽与信道带宽。信号带宽由信号频谱密度或功率频谱密度在频域的分布规律决定，同样的消息采用不同的调制方式传输时，所需的频带宽度不同；系统带宽由电路系统的传输特性决定；信道带宽由信道的传输特性决定。其中，信号带宽越小，有效性则越好，而系统带宽和信道带宽越大，则有效性越好。

比如，就信号带宽来说，调幅信号有效性优于调频信号；就信道带宽来说，同轴电缆、光纤传输介质比电话线带宽大，传输能力强。

2. 可靠性指标

模拟通信系统的可靠性指标常用信噪比或均方误差来衡量。信噪比是指接收端信号S的平均功率与噪声的平均功率N之比，即S/N。在相同的条件下，系统输出端的信噪比越大，则系统抗干扰能力越强，可靠性越好。在同样的信道信噪比条件下，不同调制方式所得到的最终解调后的信噪比是不同的，调频系统的输出信噪比大于调幅系统，故可靠性比调幅系统的好。

通信系统的有效性与可靠性是一对矛盾。比如，调频信号所需传输带宽大于调幅信号，但其抗干扰能力比调幅信号好。也就是说，调频信号和调幅信号相比，可靠性更好，但有效性更差。

3.1.4.2 数字通信系统的性能指标

1. 有效性指标

数字通信系统的有效性常用码元传输速率或信息传输速率来衡量。

(1) 码元传输速率

码元传输速率通常也称码元速率、数码率、传码率、码率、信号速率或波形速率。码元传输速率是指单位时间（每秒）内传输的码元数目，单位为波特（Baud），常用符号“B”表示。例如，某通信系统每秒传送4800个码元，则该系统的码元传输速率为4800 B。

(2) 信息传输速率

信息传输速率也称信息速率、传信率或比特率等。信息传输速率是指单位时间（每秒）内传送的信息量，单位为比特/秒（bit/s），简记为b/s或bps。例如，若某一信息源每秒传送1200个符号，而每一符号的平均信息量为1 bit，则该信息源的信息传输速率为1200 b/s。

2. 可靠性指标

数字通信系统的可靠性常用误码率或误比特率来衡量。

(1) 误码率P_e

$$P_e=\frac{\text{单位时间内系统传错的码元数}}{\text{单位时间内系统传输的总码元数}}$$

(2) 误比特率P_b

$$P_b=\frac{\text{单位时间内系统传错的比特数}}{\text{单位时间内系统传输的总比特数}}$$

3.2 信息传输技术

信息传输技术涵盖的内容非常广泛，可从不同的角度和侧面对其进行描述。本节依据传输信道形式和传输方式的不同，将信息传输技术分为光纤传输技术、微波传输技术和卫星传输技术三大类。

3.2.1 光纤传输技术

信息社会对信息传输带宽的要求越来越高，需要一种高速率、高带宽的传输技术，光纤传输技术无疑是目前最适应这种要求的传输技术。所谓光纤传输技术，是指将要传送的语音、图像和数据信号等调制在光载波上，以光纤作为传输介质的通信技术。光纤传输技术已成为包括军用、民用在内的各种高速通信网的重要支柱。为什么光纤能够传输信息，它又是如何传输信息的呢？

3.2.1.1 光纤传输原理

光纤通常是用非常透明的石英玻璃拉成的细丝。目前实用的光纤绝大多数采用由“纤芯”和“包层”两个圆筒组成的结构形式。在制作光纤时，必须使纤芯材料的折射率大于包层的折射率。原因如下：光学知识告诉我们，光线从高折射率的介质射向低折射率的介质时，其折射角将大于入射角，如果入射角足够大，光线就会出现全反射，即光线遇到包层时就会折回纤芯（见图3.4）。这个过程不断重复，光就沿着光纤传输下去。

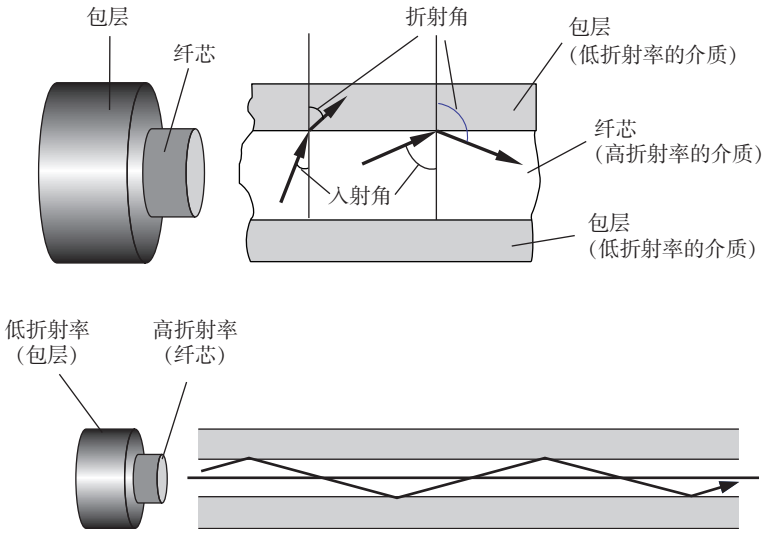


图3.4　光纤传输原理示意图

图3.4中只画了一条光线，实际上，只要从纤芯中射到纤芯表面的光线的入射角大于某个临界角度，就可以产生全反射。因此，可以存在许多条不同入射角度的光线在一条光纤中传输，这种光纤就称为多模光纤（见图3.5）。但是，光脉冲在多模光纤中传输时会逐渐展宽，造成失真。因此多模光纤只适合近距离传输。若光纤的直径减小到只有一个光波的波长，则光纤就像一根波导那样，它可使光线一直向前传播，而不会产生多次反射。这样的光纤就称为单模光纤（见图3.5）。单模光纤的纤芯很细，直径只有几微米，制造成本较高。但单模光纤的衰耗较小，在2.5 Gbps的高速率下可传输数十公里而不必采用中继器。

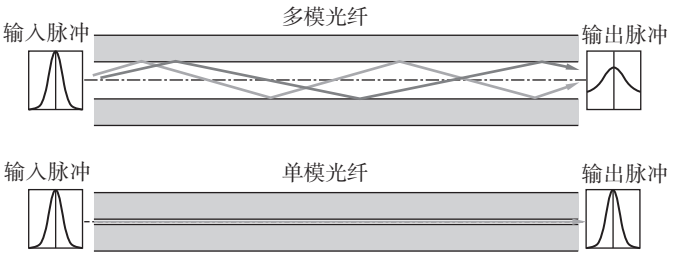


图3.5　多模光纤和单模光纤示意图

由于光纤非常细，因此光纤在实际应用时必须用适当的方式加工成光缆。光缆一般由多根（例如4根、6根，甚至上百根）光纤和辅助联络信号线、加强构件以及外护层等组成。这样处理后，可使光缆的机械强度大大提高，从而满足工程施工的强度要求。图3.6所示为8芯光缆剖面示意图，中心为钢加强芯，8根紧套光纤和2根铜信号线围着带包皮的钢加强芯缠绕。

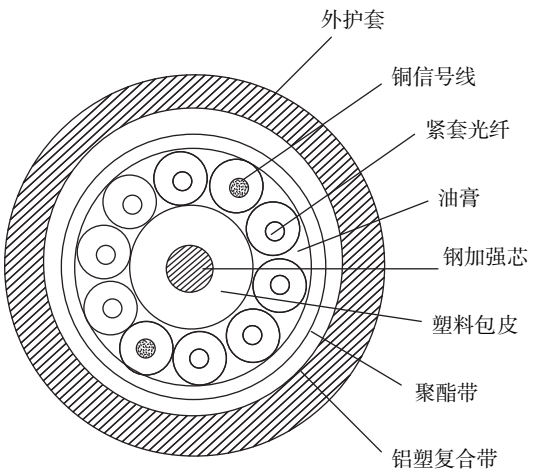


图3.6　8芯光缆剖面示意图

3.2.1.2 光纤通信系统

按传输信号的种类可将光纤通信系统分为光纤模拟通信系统和光纤数字通信系统。下面以某光纤数字通信系统为例介绍其组成和基本原理。

1. 光纤通信系统的构成

典型的光纤数字通信系统主要由电端机、光端机、光缆线路、光中继器、备用系统和辅助系统组成（见图3.7）。

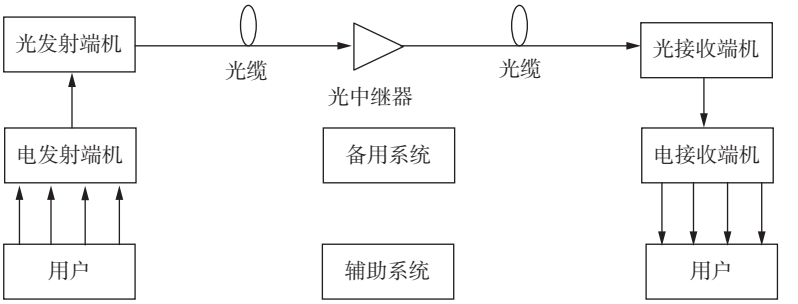


图3.7 光纤通信系统的构成

2. 各组成部分的作用

(1) 电发射端机。电发射端机包括PCM基群终端机和高次群复接（或分路）设备。电发射端机的作用是把模拟信号转换为数字信号，完成PCM编码，并按照时分复用的方式把多路信号复接、合群，从而输出高比特率的数字信号。

(2) 光发送端机。光发送端机主要完成电光转换，包括输入接口、线路编码器、调制电路、光源及控制电路等部分。

(3) 光中继器。在长途光纤通信系统中，由于光纤存在的损耗会造成光能量的损失，光纤的色散会造成光脉冲的畸变，从而引起系统性能劣化，使信息传输质量下降。因此，每隔一定距离必须设置一个光中继器（或称为光再生器），以补偿由光纤传输所产生的信号衰减和畸变，使光信号得到再生。

(4) 光接收端机。光接收端机的作用是将通过光纤传来的光信号变为电信号，再经放大、均衡和定时再生，恢复为原数字信号。

(5) 电接收端机。电接收端机完成与电发射端机相反的转换。

(6) 备用系统。由于光器件的可靠性比电子器件差，为了保证通信系统的畅通，必须设置备用系统。备用系统包括光端机、光纤和中继器，但电端机一般无备用系统。备用系统可供一个或多个主用系统共用，当某一个主用系统出现故障时，即可切换至备用系统。

(7) 辅助系统。辅助系统是为了保证光纤通信系统的信号可靠传递，以及整个光纤通信网的管理、运行和维护而设置的，主要包括监控管理系统、公务通信系统、自动切换系统、告警处理系统和电源供给系统等。

3.2.1.3 光纤传输技术的特点

光纤传输技术有如下优点。

(1) 传输频带非常宽，通信容量大。目前使用的光波频率比微波高10^3~10^4倍，通信容量大约可增加10^3~10^4倍。理论上，两根光纤可传送上百万个电话和上百套电视节目。这一特点

使得部队的情报中心可以远离雷达站、无线发射台等电磁波发射源，又能拥有高速通信通路，其好处在于既确保了安全，又能保证指挥中心与前沿部队间的通信需要。

（2）传输损耗小，中继距离长，对远距离传输特别经济。光纤通信的中继传输距离比金属材料的线缆远得多，可以达到100 km以上，是同轴电缆的几十倍。

（3）抗雷电和电磁干扰性能好。这在有大电流脉冲干扰的环境下尤为重要。光纤是非金属的光导纤维，即使工作在强电磁环境或者处于核爆炸后的强大电磁干扰环境中，光纤也不会产生感应电压与电流。这使得光纤可以靠近高压输电线和电气化铁路进行铺设，有利于在多雷地区、飞机上以及保密性较高的军政单位使用。

（4）无串音干扰，保密性好，不易被窃听或截取数据。光信息是被限制在光纤内传输的，不会逸出光纤，所以光缆的光纤之间不会串音，也不易被窃听。

（5）体积小，重量轻。这在现有电缆管道已拥挤不堪的情况下特别有利。光纤直径一般只有几微米到上百微米数量级，相同容量话路光缆要比电缆轻90%~95%，直径不到电缆的1/5。

（6）资源丰富，节约有色金属和能源。光纤的纤芯和包层的主要原材料是二氧化硅，资源丰富且价格便宜，取之不尽，用之不竭。而电缆所需要的铜铝等矿产资源是有限的。

3.2.1.4　光纤传输技术的应用

光纤通信主要应用于三个层次：干线传输、中继传输和接入传输。

光纤通信首先应用于市话局之间，构成光纤本地网；之后在长途通信中使用，构成全国性的光纤网，成为信息基础设施中的物理传输介质；接着又发展到海底光缆通信系统，用于越洋通信或短距离越岛、沿海岸等通信，其中比较著名的有横跨大西洋和太平洋的海底光缆通信系统。

在军事领域，光纤传输技术以其高速、宽带、高质量的特性已成为战场信息网络的重要传输链路。则如，野战光缆及其设备可作为野战综合通信系统的干线节点之间、固定用户群与节点之间的一种连接链路；也可作为指挥、控制、通信、情报系统之间的重要传输链路，实现指挥、控制、通信及情报系统的各个子系统与雷达、战场监视、定位火控指挥系统、防空系统、气象报知系统的互通；还可用于战场战术、战役信息系统与国防通信网乃至全球信息网络的宽带互连线路。目前，我军的军用电话自动交换网和军事综合信息网均以光缆传输网为主要承载方式。

3.2.2　微波传输技术

微波传输技术是在微波频段利用视距和中继的方法传输信息的一种无线传输技术。微波传输技术可以实现大容量通信，它不仅能够传输大容量的数据与语音业务，而且能传送频带很宽的彩色图像信号，是应用十分广泛的一种传输方式。

3.2.2.1　微波传输原理

无线电波传输是信息传输中的一种非常重要的方式，适合于空间传输的无线电波是一种电磁波，其传输速度等于光速。无线电波按频率或波长可分段命名，具体划分和命名如表3.2所示。

表3.2　无线电频率划分

波段名称	波长范围	频率名称	频率范围	代号
长波	1~10 km	低频	30~300 kHz	LF
中波	100~1000 m	中频	300~3000 kHz	MF
短波	10~100 m	高频	3~30 MHz	HF

（续表）

波段名称		波长范围	频率名称	频率范围	代号
超短波		1~10 m	甚高频	30~300 MHz	VHF
微波	分米波	1~10 dm	特高频	300~3000 MHz	UHF
	厘米波	1~10 cm	超高频	3~30 GHz	SHF
	毫米波	1~10 mm	极高频	30~300 GHz	EHF

各波段的传播特性是不同的。例如，中波主要沿地面传播，绕射能力比较强，适合广播和海上通信；短波具有较强的电离层反射能力，适合于环球通信；超短波和微波以直线传播为主，因此可用于视距或超视距中继通信。

微波的传播特性如同光波，是一种“视距”传播。这种工作方式要求天线具有强方向性，并具有足够的架设高度，信号在传播中受到的主要影响是视距传播中的直射波与地面反射波之间的干涉。与利用电离层反射进行的“超视距”传播相比，视距微波的传播特性稳定，受外界干扰也比较小。

在微波接力通信中，电波是在低层大气层传播的，会受到反射、折射、散射等的影响，从而使接收点的场强产生附加的损耗。现设球面半径为R_0，通信两端A和B的天线高度分别为h_1和h_2，最大视线距离为d，当天线高度h_1和h_2给定时，最大视距为

$$d = \sqrt{2R_0}\left(\sqrt{h_1} + \sqrt{h_2}\right)$$

可见，天线越高，视距就越大。在平原地带可利用铁塔或高层建筑物为基础，以提高天线的高度；而在山区则可借助山峰架设天线，这样可增长视线距离。

微波的视线距离通常为50 km左右。若采用100 m高的天线塔，则传播距离可增大到100 km。为了实现远距离的通信，就需要在两个终端站之间建立若干个中继站，以接力的方式逐站依次转发传递信号（见图3.8）。这种利用地面中继站转发信号的超视距、多路无线电通信系统，称为微波中继传输系统。

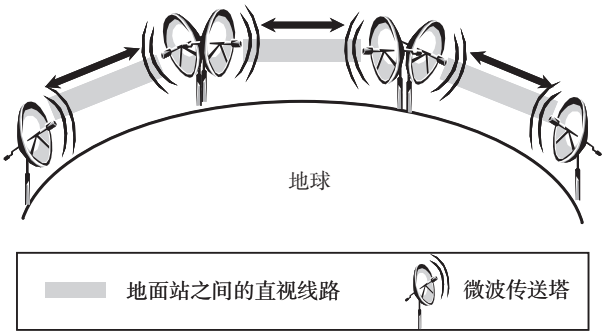


图3.8 微波接力示意图

3.2.2.2 微波通信系统

微波中继传输系统的组成如图3.9所示。

按照工作性质与任务的不同，线路上的微波站分为三类：微波端站、微波中继站和微波分路站。

微波端站是指处于线路两端的微波站，其主要功能是：发信时，由数字终端机把各用户信号转换为时分多路信号，经信道中的发信机调制变频后，通过天线发射出去；收信时，由天线接收到的微波信号经收信机解调，再经数字终端机分路，分出各用户信号。

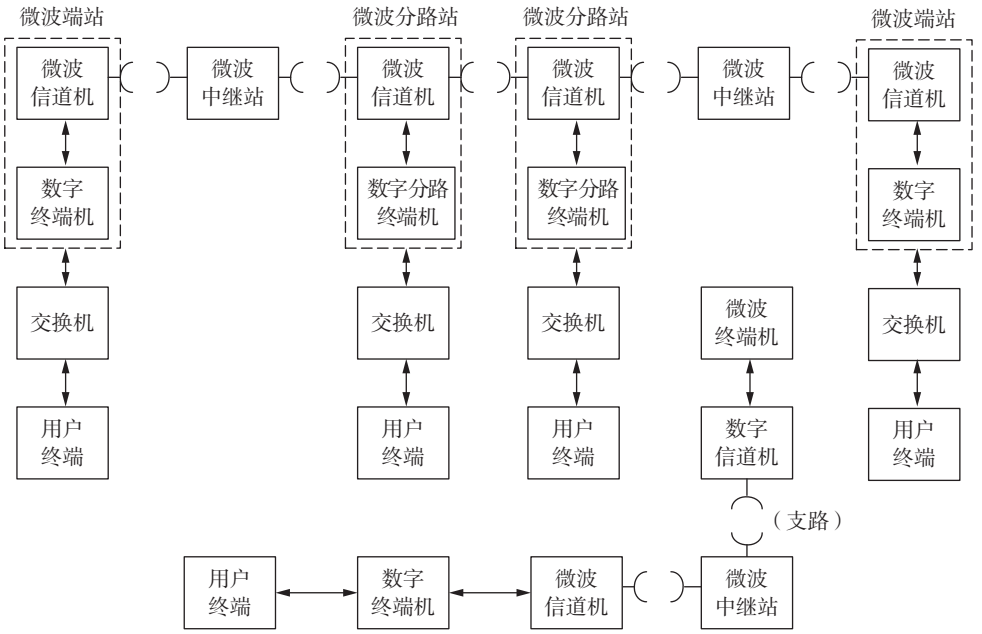


图3.9　微波中继传输系统方框图

微波中继站是线路中间只完成信号转接而不设置上、下话路的微波站。通常采用基带转接方式，即对接收的微波信号经下变频与解调，得到数字基带信号，对其整形再生后，再上变频发射出去。采用这种转接方式，可在中继站消除信号的干扰和畸变，从而避免失真和噪声逐站积累。

微波分路站是线路中间需要设置上、下话路的微波站，它将天线接收到的微波信号通过微波收信机进行下变频与解调，经数字分路终端机从中分出一部分话路，并可插入一部分新话路，再经对另一方面的微波发信机进行调制与上变频，由天线发射出去。微波分路站既能转发又能分出或插入话路，分出的话路可通过交换机接至用户或数字微波接力通信支线。

用户终端指直接由用户使用的终端设备机、计算机、调度电话机等。

交换机是用于功能单元、信道或电路的暂时组合，以保证按要求进行通信操作的设备。用户可通过交换机进行呼叫连接，建立暂时的通信信道或电路。这种交换既可以是模拟交换，也可以是数字交换。

数字终端机的基本功能是把来自交换机的多路音频模拟信号转换成时分多路数字信号，送往数字微波传输信道，以及把数字微波传输信道收到的时分多路数字信号转换成多路模拟信号，送到交换机。

3.2.2.3　微波传输技术的特点

微波传输技术有一些显著的特点，其优点主要包括以下几个方面。

(1) 工作于微波波段，频带很宽，相当于超短波以下所有各波段宽度总和的10 000倍，可以容纳许多互不干扰的宽频段通信电路。

(2) 微波的波长很短，可以在合理的尺寸下，做出高增益、方向性强的天线。这不仅可以大大减小发射功率，而且也减少了遭受各种干扰和泄密的可能性。微波天线发出的电磁波束较窄，不易被侦察和截获。

(3) 由于诸如工业干扰、天电干扰等各种干扰的主要频谱成分处在比微波低得多的波段，所以对微波通信的影响比较小，故而微波通信的质量较高。此外，微波的传播受季节、昼夜变化的影响较小，信道较为稳定，可视为恒参信道。即使是在磁暴、核爆炸的情况下，线路仍能保持畅通，所以稳定性和可靠性较高。

微波传输技术的缺点如下。

(1) 相邻站之间必须是通视的，不能有任何障碍物。

(2) 通信距离受天线高度和接力站数目限制。

(3) 与有线通信相比较，保密性较差。

(4) 电波在空间传播时，由于大气和地形地物多径反射的影响，会引起衰落现象，因而通信质量和稳定性仍然存在一些问题。

3.2.2.4 微波传输技术的应用

微波传输技术是地面无线宽带的主要传输手段之一，可用于干线通信、专用通信、移动通信的中继线路等场合。

在军事领域，微波传输技术不仅能用于战术通信，也能用于战略通信。它可实现各总部和各军兵种之间的干线通信、边海防通信、抢险救灾通信，可作为卫星地面传输线路和野战地域通信网的传输干线。

3.2.3 卫星通信技术

1. 卫星通信概念

卫星通信是指利用人造地球卫星作为中继站转发无线电信号，在两个或多个地球站（包括地面、水面和低层大气中）之间进行的通信。通信卫星的作用相当于距地面很高的中继站。由于卫星通信的无线电信号频率是处于300 MHz~300 GHz的微波波段，所以卫星通信实际上是利用卫星作为中继站的微波中继通信方式。

用户利用卫星进行通信的基本过程是：用户信号经地面通信线路（光缆、微波接力、电缆等）送到地球站，地球站发信设备对信号进行处理，变为上行微波信号，进行功率放大后经由天线发往卫星；信号经上行路线传播到卫星，卫星上的通信天线把收到的微弱信号送给卫星转发器，进行放大处理后变为下行微波信号，再经天线转发并经下行路线传播到对方用户所在区域的地球站。该地球站对卫星转发下来的微弱信号进行放大处理，之后由地面通信线路传送给有关用户。

卫星通信系统的组成包括通信卫星分系统（也称为空间分系统）、地球站分系统、跟踪遥测及指令分系统，以及监控管理分系统四大部分（见图3.10）。前两个分系统主要用于通信，后两个分系统一般起支持和保障作用。

通信卫星分系统的主体是信号转发系统，此外还包括星上遥测及指令系统、控制系统和电源系统等。信号的转发是靠卫星上的转发器（微波收发信机）和天线来实现的。一个卫星的信号转发系统可能有一个或多个转发器，每个转发器有一定的工作带宽，转发器的带宽和数量决定了卫星通信系统的容量。

地球站分系统就是微波收发信台站，用户通过它们接入卫星通信线路。通常，地球站包括地面站、机载站、舰载站等多种类型。典型的地球站由天线、跟踪和伺服设备、发射设备、接收设备、信道终端设备、保密终端设备、供电电源等组成。

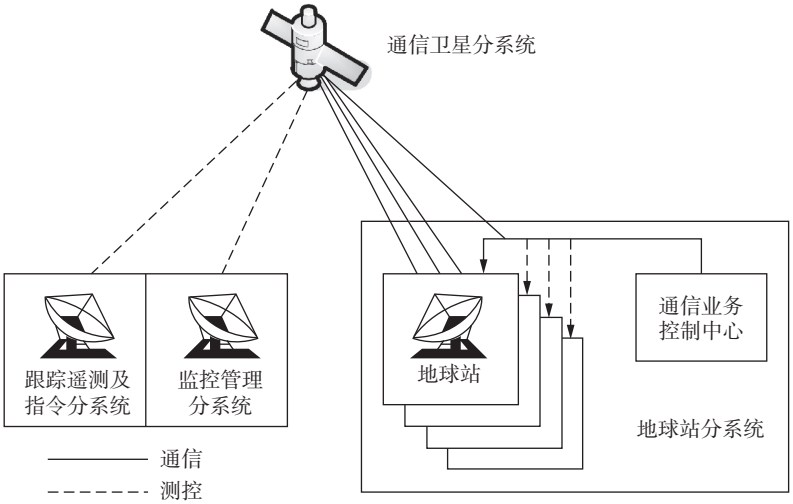


图3.10　卫星通信系统示意图

跟踪遥测及指令分系统主要完成对卫星进行跟踪测量的任务，控制其准确进入静止轨道上的指定位置。卫星正常运行后，要定期对卫星轨道进行修正。

在通信业务开通前后，由监控管理分系统进行通信性能的监测和控制，例如对卫星转发器功率、卫星天线增益以及各地面站发射的功率及带宽等参数进行监控，从而保证系统正常通信。

2. 卫星通信链路与频段

（1）卫星通信链路

在传统卫星通信中，通常使用上行链路（Uplink）和下行链路（Downlink）来表示卫星与地球站之间的传播路径，其中上行链路为地球站发送信号到卫星所经过的通信路径，下行链路为卫星发送信号到地球站所经过的通信路径。

在卫星移动通信中，对于通信链路通常划分为下列4类。

- 前向链路（Forward Link）：从信关站（Gateway）到移动站方向的链路。
- 后（反）向链路（Return Link）：从移动站到信关站方向的链路。
- 用户链路（Subscriber Link）：移动站与卫星之间的链路。
- 馈电链路（Feeder Link）：信关站与卫星之间的链路。

上述链路均属于卫星与地球站之间的通信链路，统称为星地链路。与此对应，卫星之间用于实现信息交换的通信链路称为星际链路。

星际链路（Inter-Satellite Link，ISL）又称为星间链路，有时也称为交叉链路（Cross link）。

（2）卫星通信工作频段

为保证相互之间不干扰，卫星通信的工作频率是不可以随意使用的，ITU主持召开相关的会议来负责频率的分配和协调。

国际上，频率是按区域来划分的。全球分为3个区域：Ⅰ区包括欧洲、非洲、前苏联的亚洲部分、蒙古、伊朗西部边界以西的亚洲国家；Ⅱ区包括南美洲、北美洲、格陵兰岛、夏威夷群岛；Ⅲ区包括亚洲的其他部分、大洋洲。

除此之外，在卫星通信中，还按业务类型划分工作频率。在这3个区域内，卫星频带分别被分配给各种卫星业务。ITU规定的卫星业务种类如下：

- 卫星固定业务
- 卫星移动业务
- 卫星无线电导航业务
- 卫星无线电定位业务
- 卫星广播业务
- 卫星气象业务
- 业余卫星业务

上述业务中，卫星固定业务和卫星移动业务属于卫星通信的范畴，卫星广播业务有时也划入通信的范围。对于一种给定的业务，在不同的区域既可以分配给不同的频带以防止同频干扰，也可以分配给相同的频带以提高频谱利用率。表3.3给出了卫星通信系统可以使用的各类频率及其频段名称。

表3.3 卫星通信系统使用的频率

频率范围（GHz）	波段代码	ITU的频段表示方法		
		名称	米制划分	缩写
0.03~0.3		甚高频	米波	VHF
0.3~1.0		特高频	分米波	UHF
1.0~2.0	L			
2.0~3.0	S			
3.0~4.0		超高频	厘米波	SHF
4.0~8.0	C			
8.0~12.0	X			
12.0~18.0	Ku			
18.0~27.0	K			
27.0~30.0	Ka			
30.0~40.0		极高频	毫米波	EHF
40.0~75.0	V			
75.0~110.0	W			
110.0~300.0	mm			
300.0~3000.0	μm	至高频	微米波	THF

目前，大部分通信卫星尤其是商业通信卫星使用C波段和Ku波段，C波段的上行链路为5.625~ 6.425 GHz，下行链路为3.4~4.2 GHz，转发器带宽可达800 MHz；Ku频段的上行链路为14~14.5 GHz，下行链路为11.7~12.2 GHz、10.95~11.2 GHz或11.45~11.7 GHz。国内和区域性通信卫星多数也使用该波段。许多国家的政府和军事卫星用X波段，上行链路为7.8~8.4 GHz，下行链路为7.25~7.75 GHz，这样与民用卫星通信系统在波段上分开，避免互相干扰。

随着C波段和Ku波段越来越拥挤，不仅对原有的波段进行了扩展，而且开始使用20~30 GHz的Ka波段，其上行链路频率为27.5~31 GHz，下行链路频率为17.7~21.2 GHz。用该波段时的可用带宽可增大到3.5 GHz，因此有很大吸引力，但受降雨的影响相当严重。

3. 卫星通信中的多址接入技术

卫星通信多址接入方式是指卫星通信系统内的多个地面站以何种方式接入卫星并从卫星接收信号。

多址接入方式首先要解决的问题是尽可能灵活地把网络内的所有用户通过卫星互连起来，同时要尽量有效地利用卫星的频率和功率资源。另外，还需要考虑到系统对业务类型和网络扩容的灵活性、对不同业务类型的适应能力、经济性、地球站的复杂度、系统的安全保密性等问题。

实现多址接入的技术基础是信号分割，也就是在发送端要进行恰当的信号设计，使系统中各地球站所发射的信号有所差别；而各地球站的接收端则具有信号识别能力，能从混合着的信号中选择出本站所需的信号。

目前，卫星通信系统中常用的多址接入方式有FDMA、TDMA和CDMA及其混合形式。

（1）FDMA方式

FDMA的基本特征是，把卫星转发器的可用射频频带分割成若干互不重叠的部分，分配给各地球站的各载波使用。因此，FDMA信号以频率为参量来进行分割，而在时间和空间上并不分割。所有用户终端的发射信号在频率上是互不重叠的，接收端根据频率来接收属于自己的信号，图3.11给出FDMA利用频率实现信号分割的示意图。

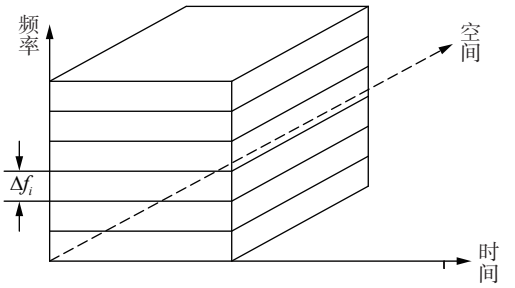


图3.11　FDMA利用频率分割信号的示意图

在FDMA系统中，每个载波都是相对独立的，可以采用独立的调制方式、基带信号形式、编码方式、信息速率、占用带宽等，而不必考虑其他载波采用什么方式，只要它们在频谱上不与本载波重叠即可。

根据每个地球站在其发送载波中是否采用复用技术，通常将FDMA分为两大类：每载波多路信道（MCPC）和每载波单路信道（SCPC）。另外，在多波束环境中，通常采用卫星交换的FDMA（SS-FDMA）以实现不同波束区内地球站之间的相互通信。

MCPC-FDMA方式主要用于业务量比较大、通信对象相对固定的点–点（点–多点）干线通信，如IDR业务；而SCPC-FDMA方式较适合稀路由应用环境（站较多，每站业务量小）。

FDMA方式的优点是技术成熟，设备简单，每个载波可以采用独立的调制、编码方式；主要缺点是：由于转发器的非线性，多载波工作时会产生互调噪声，存在大载波抑制小载波的现象；为减小互调，要求转发器有输出补偿（回退）；需要设置保护带以避免邻道干扰；对于MCPC-FDMA方式，信道分配不灵活，业务较闲时频带利用较低。

（2）TDMA方式

TDMA的基本特征是，把卫星转发器的工作时间分割成周期性的互不重叠的时隙（每个时隙也称为分帧，一个周期则称为一帧），分配给各站使用，各站信号在频率和空间上可以重叠（见图3.12）。

在TDMA系统中，由于所有上行链路地球站的发射载波频率是相同的，所以系统必须让所有地球站在时间上同步，以使每个站都只在指定时间段内发射，而不会因为误入其他时间段而造成相邻站之间的相互干扰。此卫星和所有地球站之间的时间同步称为网络同步。对于接收站来说，也需要网络同步，以便在一个特定时隙内接收某给定地球站发送的信号。

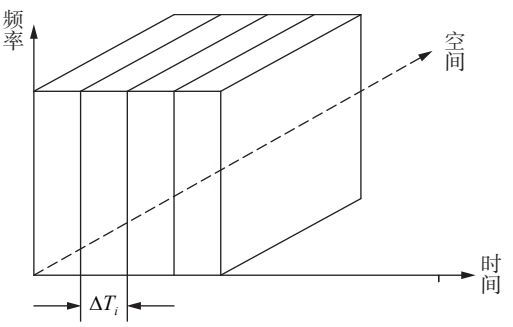


图3.12　TDMA利用时间分割信号的示意图

在TDMA方式中，某个时刻转发器（或某一频率段）中只有一条TDMA载波，每个上行链路地球站被分配在一个预先规定好的时间段内发送，在该时间段内，卫星的功率和频率资源均由该地球站发射的上行链路载波使用。由于没有其他载波在该时隙内同时使用卫星，因此不存在互调和大载波抑制小载波的现象，卫星的功率放大器可以工作在饱和区，从而能得到最大的卫星输出功率。但对于系统中存在的站数较多、各站业务量不大、各站链路质量差距较大的应用环境（如卫星移动通信）以及多波束应用环境，传统TDMA方式存在明显的不足，为此，又发展出了卫星交换TDMA（SS-TDMA）和多载波TDMA，尤其是后者，近几年在VSAT卫星通信领域得到了广泛的应用。

MC-TDMA方式是指在一个TDMA系统中采用多条信道速率相对较低（小到几十Kbps，高到20 Mbps）的载波，每条载波以TDMA方式工作。对于地球站来说，虽然系统中同时有多条TDMA载波，但某个时候每个站只能在一条TDMA载波上发送或接收；如果要同时在两条载波上发送或接收，则需配备两套设备。

（3）CDMA方式

CDMA的基本特征是各站所发的信号以码形参量来进行分割，各站采用各不相同的、相互正交或准正交的地址码分别调制各自要发送的信号，而在频率、时间和空间上不分割（见图3.13）。

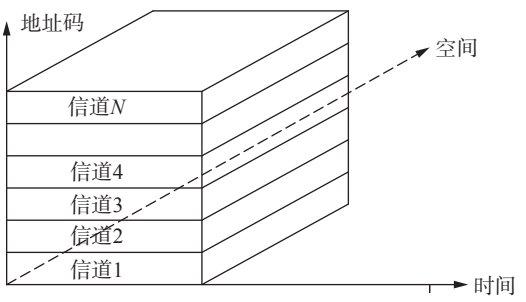


图3.13　CDMA利用地址码分割信号的示意图

CDMA方式的工作原理如下。利用自相关特性非常强而互相关特性比较弱的周期性码序列作为地址信息（称为地址码），对被用户信息调制过的载波进行再次调制，使其频谱大为展宽（称为扩频调制）；经卫星信道传输后，在接收端以本地产生的已知地址码为参考，根据相关性的差异对接收到的所有信号进行鉴别，从中将地址码与本地地址码完全一致的宽带信号还原为窄带而选出，其他与本地地址码无关的信号则仍保持或扩展为宽带信号而被滤除（称为相关检测或扩频解调）。

CDMA有两种实现方式。一种是直接序列扩频（DS）方式，用地址码直接对信号进行调制来得到扩频信号；另一种是跳频扩频（FH）方式，用地址码控制频率合成器，使它产生出能在较大范围内周期性跳变的本振信号，再用它来与已调信号载波进行混频来得到扩频信号。

CDMA的主要优点包括：宽带传输，抗多径衰落性能较好；信号频谱的扩展和相关接收具有较好的信号隐蔽性和保护性，抗干扰能力也较强；具有扩频增益，允许相邻波束使用相同频率，频率复用能力强；能充分利用语音激活来提高容量；移动通信中具有软切换功能；具有软容量，增加用户只会影响性能。

当然，CDMA也存在一些缺点，主要有：要占用很宽的频带，频带利用率一般较低；选择数量足够的可用地址码组较为困难；接收时，对地址码的捕获与同步有一定时间；需要进行严格的功率控制；受扩频码片速率的限制，主要用于低速业务。

CDMA方式在移动通信中有广泛的应用，尤其适合军事卫星通信系统及多波束卫星移动通信系统。

4. 卫星通信中的调制技术

为使数字信号在带通信道中传输，必须对数字信号进行调制处理。调制方式的选择与所用的信道有密切关系。卫星信道的主要特点是功率受限（有时也会频带受限），同时可能还具有非

线性特性、衰落特性和多普勒频移。功率受限主要是由于卫星发射机的EIRP相对较小，传输损耗很大，但接收终端的天线增益相对较小，导致解调器输入端的信噪比很低，通常其E_b/N_0的值低于10 dB，远远低于有线/光纤通信系统以及蜂窝移动通信系统的E_b/N_0值（通常在30~40 dB）。卫星通信信道的非线性来自于高功率放大器，为了充分利用发射机的功率，其行波管放大器或固态功率放大器常常工作在非线性的饱和区。

当前研究结果表明，BPSK、QPSK、OQPSK、π/4-DQPSK、MSK和GMSK等都可以用于卫星通信；而信道非线性、干扰、多径效应和多普勒频移的影响以及实现均衡的困难决定了高阶调制方式（如8PSK、64QAM、OFDM等）不宜用于卫星通信信道。

以PSK调制为例。它是用载波的相位携带信息的。MPSK信号的载波相位共有M个可能的取值，每一个载波相位对应着M个元素的符号集中的一个符号，在某一个符号间隔内载波的相位取该符号对应的相位值。调制的过程就是将待传输的符号转换为载波的相位，而解调的过程则是将载波的相位转换为所传输的符号。通常待传输的信息流是二进制比特流，这时取$M = 2n$（$n = 1, 2, 3\cdots$），即每n个二进制比特对应于一个符号。因此在调制时还需要将二进制比特流转换为相应的符号，解调时再还原为二进制比特。如果$M=2$，则比特与符号是一致的，不需要转换。

MPSK信号的数学表达式如下：

$$s(t) = \mathrm{Re}\left\{u(t)\exp \mathrm{j}(2\pi f_c t + \varphi_k + \lambda)\right\},\ (k-1)T_0 \leqslant t \leqslant kT_0 \tag{3.1}$$

其中$u(t)$是基带信号，f_c是载波频率，$\varphi_k \in \{0, 2\pi/M, 4\pi/M, \cdots, 2(M-1)\pi/M\}$，$\lambda$是载波的初始相位，$T_0$是符号间隔。

当$M = 2$时就是BPSK信号，$M = 4$时就是QPSK信号。随着M的增加，已调信号的频谱效率增加，而功率效率则下降。

5. 卫星通信中的信源编码技术

信源编码也称数据压缩，它是将信源输出信号有效地映射成符号序列的过程，提高了信息传输的有效性，可进一步分为语音编码和图像编码等。任意给定的信源都有一个表征其不确定性的称为“熵”的量，它是无失真数据压缩的下限。信源编码定理指出：在允许一定的失真情况下，存在最小数量的比特，描述独立同分布的信源输出。

（1）语音编码技术

语音编码是将待传输的模拟语音信号转换成数字语音信号，使其能够在数字通信系统中传输。

语音的压缩编码方法归纳起来可以分为3大类：波形编码、参数编码和混合编码。

常用语音编码方法有脉冲编码调制（PCM）、增量调制（DM）、自适应差分脉冲编码调制（ADPCM）、线性预测编码（LPC）、规则脉冲激励线性预测声码器（RPE-LPC）、码激励线性预测声码器（CELP）、矢量和激励线性预测声码器（VSELP）、短延时码激励线性预测声码器（LD-CELP）和多带激励声码器（MBE）等。

（2）图像编码技术

图像编码是将需要传输的图像按一定的算法进行编码处理，去除冗余部分，用尽量低的速率传输尽量高质量的图像。

目前，针对不同的图像类型主要存在下列3类图像编码标准。

1）静止图像压缩编码（Joint Photographic Experts Group，JPEG）标准

JPEG是适用于彩色和单色多灰度或连续色调静止数字图像的压缩标准。基本JPEG算法先将图形分割成8×8的二维块，再以块为单位采用离散余弦转换（DCT）进行无损压缩；然后对DCT系数进行量化，对量化后的DCT系数进行哈夫曼编码，使其熵达到最小，可达80∶1的压缩效果。

2）运动图像压缩编码（MovingPicturesExpertsGroup，MPEG）标准

MPEG标准是针对运动图像而设计的，平均压缩比可达50∶1，有统一的格式，兼容性好。

3）视频会议图像压缩编码标准——H.26X标准

ITU-T针对视频会议系统制定了相应的视频会议图像压缩编码标准，主要有如下几项。

- H.261：用于可视电话/电视会议的建议标准，压缩比可达50∶1左右。
- H.263：是支持传输速率低于64 Kbit/s的窄带信道的视频编码。
- H.264：不仅针对视频会议系统，而且涵盖了电视广播、网络流媒体、多媒体信息的数字存储、数字影院等各方面的应用。H.264的良好网络适应性和内在的抗丢包能力、抗误码机制，使它不仅适于IP传输方式，也适合丢包严重、时延和抖动复杂的无线信道。与前述图像编码标准相比，H.264的编码效率更高、图像质量更好。

6. 卫星通信中的信道编码技术

信道编码又称差错控制编码，根据一定的规则，适当增加信号的冗余度，使其具有一定的检测错误和纠正错误的能力。与信源编码的数据压缩相反，信道编码技术通过人为地按照一定规则增加冗余，以克服信息传递过程中受到的噪声和干扰的影响，使恢复的信源信息的错误概率尽可能小，从而提高信息传输的可靠性。

按照信道编码的不同功能，可以将其分为检错码、纠错码和纠删码。检错码仅能检测误码；纠错码仅可纠正误码；纠删码则兼有纠错和检错能力，当发现不可纠正的错误时能够发出错误指示，或者简单地删除出现不可纠正错误的信息段落。

对于纠错码，通常使用编码增益来衡量其纠错能力。编码增益定义为在给定编码码率和调制方式的情况下，为获得相同的误比特率，未使用编码时所需要的信噪比与采用编码后所需要的信噪比的对数差值。

针对这三种不同的信道编码，相应地有三种差错控制方式。

（1）自动重发请求（ARQ）方式。发送端发送只能检错的码字，接收端收到后根据编码规则检测传输过程中是否有错。如果有错，则通过反向信道通知发送端重发。重发的次数可能是一次，也可能是多次，直到接收端认为传输无错为止。ARQ方式的优点是工作原理简单易于实现，缺点是有延迟。因此主要用于对实时传输要求不高的数据传输系统。

（2）前向纠错（FEC）方式。发送端发送能够纠错的码字，接收端收到后根据编码规则进行译码，通过译码发现并纠正传输过程中的错误。FEC方式不需要反馈，特别适合于只能提供单向信道的场合。FEC的另一个特点是不要求检错重发，因此时延小，实时性好，可用于对实时传输要求高的信号传输系统，如语音传输系统等。

（3）混合纠错（HEC）方式。该方式是FEC方式和ARQ方式的结合。其发送端发送既能纠错又能检错的码，接收端进行纠错译码后，如果没有检测到误码，则不再要求发送端重发；如果接收端经纠错译码后仍检测出有误码，则通过反馈信道要求发送端重发。

目前，常用的信道编码方式主要有以下几种。

（1）线性分组码

采用线性分组码方式是指将信息码序列分成长度为k的码组，然后在k个信息比特的后面加上$n-k$个监督比特（也称校验比特），构成长度为n的码组，称为(n, k)码。监督比特仅与本码组的信息比特有关，与其他码组的信息比特无关。一个(n, k)分组码的编码效率（简称码率）定义为k/n。

（2）里德–所罗门码

里德–所罗门码（Reed-Soloman码，RS码）是为纠正突发错误而设计的一种编码。RS码不直接对比特进行编码，而是先将比特分组组成符号，然后将这些符号进行编码。

（3）卷积码

采用卷积方式码指是把k个信息比特编码成n个比特，但k和n通常很小，特别适合以串行方式传输信息，时延小。卷积码中编码后的n个码元不仅与本码组的k个信息元有关，而且与以前m($m \geqslant 1$)个码组的信息元也有关。因此，卷积码被表示为(n, k, m)。m称为卷积码的编码存储度。

（4）交织编码

交织编码是一种能够有效纠、检突发错误的编码方法，它改变了传输比特的顺序，使突发错码分散到几个码字中，而不是集中在一个码字中。

（5）级联码

级联码包括内码和外码这两个独立的编码，外码主要用于纠突发差错，而内码主要用于纠随机差错。当信道产生少量的随机差错时，通常内码就可以纠正误码。当产生较长的突发差错或随机错误很多，以致超过内码的纠错能力时，内码译码器产生错译，但这些错码可以通过外码译码器予以纠正。

（6）Turbo码

Turbo码是在传统级联码的基础上提出的，它采用并行卷积级联码（PCCC）软输入–软输出迭代译码结构，使编码增益大大提高，纠错性能接近香农极限。其优异的编码性能得益于两个思路的组合：（1）采用递归系统卷积码（Recursive Systematic Convolutional Code，RSC）作为构造级联的子码；（2）利用交织器将RSC进行并行级联。由于Turbo码接近随机码，有很好的距离特性，而且在编码中融入了交织码，因而有很强的纠错能力和抗衰落能力。

（7）LDPC码

LDPC（Low Density Parity Check）码，即低密度奇偶校验码，是一种线性分组码。名称来源于其校验矩阵的稀疏性，即校验矩阵中只有数量很少的非零元素，大部分都是“0”。LDPC码在高斯信道和莱斯信道下都有着良好的纠错性能，具有很多优点，例如较低的误码平台特性，可实现完全的并行操作，译码复杂度低于Turbo码，适合硬件实现，吞吐量大，具有高速译码的潜力等。LDPC码的主要缺点是，码长较大导致编译码时延较大，对于低速语音通信不太适合，也比卷积码复杂得多。

3.3　信息网络技术

信息网络技术涵盖的内容非常多，本节将主要介绍信息网络的基本概念、构成要素、常见基本结构等基本知识，以及计算机网络相关技术。

3.3.1　信息网络概念

最简单的通信是在信源和信宿这两个终端之间建立一条信息传输的通道。当信源和信宿的数量有限时，这是完全可能的。但当信源和信宿的数量较多时，比如有n对信源和信宿之间通信，则需建立$n*(n-1)/2$条固定的信息传输通道，这就会造成成本的剧增和资源的巨大浪费。尤其当终端之间数据通信量不大，信道利用率较低时，此种连接方式尤其不经济。

克服上述缺点的办法是：把所有终端连接到一个网络上，网络资源对所有终端是共享的，终端之间可以通过这些共享的网络资源传输数据。为了提高数据传输的可靠性，网络在终端之间提供多条路径。这时，每个终端只需要一个输入/输出端口，而不是$n-1$个端口。从而使终端设备大大简化，降低了整个系统的成本。这种许多通信点之间建立相互连接，而且点与点之间不只一条路径的多对多的通信系统（即多用户通信系统互联的总体）称为信息网络。

由信息网络的定义可以看出，信息网络实际上就是两台以上的通信设备经互联所构成的系统，具体来说就是一种使用交换设备和传输设备，将地理上分散的用户终端设备互联起来，实现通信和信息交换的系统。如“计算机网络”即是两台以上的计算机经互联所构成的系统。从网络的实质来看，“网络”就是交流信息、传播信息、共享信息的通道和手段。

3.3.2 信息网络的组成

信息网络的构成要素主要有通信终端设备、传输设备和节点（交换）设备。

1. 通信终端设备

通信终端设备能够将输入的信息转换为便于传输的形式，并作为网络输入设备而控制通信网的工作，起到信息处理装置与通信网之间的接口作用，主要有电话机、固定通信终端、数据通信终端（传真机、计算机）和移动通信终端等。

2. 传输设备

传输设备的作用是传输经通信终端转换而成的电信号，它由传输介质和各种通信装置构成。这些通信装置具有波形转换、调制解调、多路复用、发信和收信功能。目的是更有效地利用传输介质。

3. 节点设备

节点设备就是交换设备，实际上是一个交换机，它在通信网中主要起以下几方面的作用。

(1) 进行通信终端或中继线路之间的接续转换，连接发信终端和收信终端的交换功能。

(2) 根据通信流量状态，有效地选择通信路由，遇到故障或异常时具有通信网的网络控制和管理功能。

(3) 具有执行各种交换业务、通信业务的功能。

按不同的通信目的，节点的功能会有所不同，但上述几个主要作用都是必备的。电话通信网中的节点是电话交换机，数据通信网和图像通信网中的节点分别是数据交换机和图像交换机，计算机网络中的节点是信息处理器或路由器，有的地方又称为网关。信息网络的构成如图3.14所示。

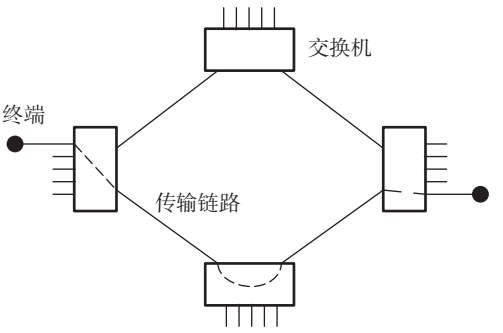


图3.14 信息网络的基本构成

3.3.3 信息网络的基本结构

信息网络的基本结构有网形、星形、复合型、环形、总线型和树形（见图3.15）。每种结构特点不同，每种网络可根据所提供的服务采用合适的拓扑结构。

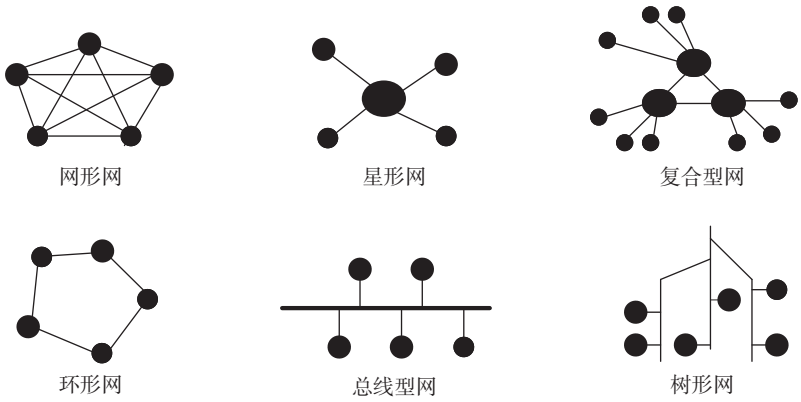


图3.15　信息网络的基本结构

1. 网形网

网形网共需$n(n-1)/2$条传输链路。优点是当某链路发生故障时，不会影响通信。缺点是n值较大时，链路利用率低，经济性较差，且网络控制分散，难度大。

2. 星形网

星形网共需$n-1$条传输链路。优点是网络控制由中心节点控制，比较简单。缺点是可靠性差，万一中心节点发生故障，则会影响通信。中心节点的维护成本较高。

3. 复合型网

复合型网是网形和星形的组合型网。在通信量较大的区域构成网形网结构，而在局部区域内构成星形网结构。这种网络结构兼取了上述两种网络的优点，较为经济合理，并有一定的可靠性。

4. 环形网

环形网是一种特殊的网形网。采取分布控制方式，在局域网中应用较多（如以太网），有单向环和双向环两种，双向环的可靠性明显高于单向环。优点是控制方式简单，费用也较低，但有时完成一次接续要经过几级转接，因而容易导致各节点交换机设备庞大。

5. 总线型网

总线型网通过总线把所有节点都连接起来，所有节点共享总线。这种结构较简单，扩展十分方便，也主要用于局域网。

6. 树形网

树形网是总线结构的一种扩展，它是一种分层网，适用于分级控制系统。与星形网相比，通信线路的总长度较短，成本较低，节点扩展灵活，但结构比星形网复杂。

3.3.4　信息网络的分类

1. 按通信内容分类

按通信内容分类，信息网络可分为电话网、电报网、电传网、移动通信网、计算机网、综合业务数字网、有线电视网等。通信网络发展初期都以电话网为主。但从发展来看，其他信息传输业务的增加速度已超过电话业务的增加速度。

2. 按通信范围分类

按通信范围分类，信息网络可分为局域网、市内网、国内网和国际网。局域网通常是网络的最小单元，它把一个较小区域内的用户相互连接起来，能在两两用户之间进行信息的交互传输。各个局域网之间如果能相互通信，就能形成一个较大的网，如市内网。

3. 按组成网络的信道分类

按组成网络的信道分类，还可以分为电缆网、短波无线电网、微波中继网、卫星网、光缆网等。这些网在传输和使用方面各有特点。例如，短波无线电网虽然容易建立，但其衰落现象严重，且容量有限；卫星网适宜远距离通信；光缆网的潜在容量很大，干扰很小。

显然，各种网是根据不同条件和需要建立的，但它们可以相互补充和备用。不过，如果要将它们组成一个网，则网间互联往往会遇到通信协议、数据格式、信令等不一致造成的困难。

3.3.5 三大信息网络

21世纪是一个以网络为核心的信息时代，网络现在已成为信息社会的命脉和发展知识经济的重要基础。通常人们所说的网络是指“三网”，即电信网络、有线电视网络和计算机网络。这三种网络对社会生活的很多方面以及对社会经济的发展已经产生了不可估量的影响。

3.3.5.1 电信网络

电信网是利用有线、无线或两者结合的电磁和光电系统，传递文字、声音、数据、图像或其他任何媒体信息的网络。具体地说，电信网包括电话网、电报网、传真网、帧中继、数字数据网（Digital Data Network，DDN）等，可为用户提供各种通信业务，如电话、传真、会议电视以及数据通信等。电信网作为人类实现远距离通信的重要基础设施，其发展经历了模拟网→数字网→同步光纤网（Synchronous Optical Network，SONET）或同步数字体系网（Synchronous Digital Hierarchy，SDH）→综合业务数字网（Integrated Services Digital Network，ISDN）的历程，其主要功能是按用户的需要传递和交流信息，以实现用户之间的远距离通信。

3.3.5.2 有线电视网

有线电视网（Community Antenna TV，CATV）是利用光缆或同轴电缆来传送广播电视信号或本地播放的电视信号的网络。CATV高效且廉价，具有频带宽、容量大、多功能、成本低、抗干扰能力强、支持多种业务、可连接千家万户的优势。有线电视网经历了以同轴电缆为主的单向电视频道节目信号传输，到以混合光纤/同轴（Hybrid Fiber Coax，HFC）网络为主的单向信号传输，再到双向HFC或采用光纤传输，从而能够提供视频点播等新业务的发展历程。

3.3.5.3 计算机网络

与用于双方交流信息的信息网络（即电信网）和用于向大众单向传播信息的传媒网络（即广播电视网）不同，计算机网络是向用户提供信息资源共享、计算能力共享的网络。鉴于现在人们的生活、工作、学习和交往都已离不开计算机网络，下面对其相关知识进行简要介绍。

1. 定义

计算机网络是利用通信设备和传输线路，将不同地理位置的功能独立的多台计算机系统互联，通过一系列协议实现资源共享和信息传递的系统。计算机网络向用户提供的最重要功能有

两个：连通性和资源共享。所谓连通性，就是计算机网络使上网用户之间都可以交换信息，好像这些用户的计算机都可以彼此连通一样。资源共享的含义则是多方面的，包括信息共享、软件共享以及硬件共享等。如计算机网络上很多主机存储了大量的电视剧和歌曲，可供上网用户有偿或无偿地读取和下载。

2. 体系结构

计算机网络由许多软硬件实体组成，需要利用标准来协调网络之间的通信。最著名的两个标准是法定的国际标准——开放系统互联基本参考模型OSI/RM（Open Systems Interconnection Reference Model，简称OSI）和国际标准TCP/IP协议栈。图3.16是TCP/IP参考模型和OSI参考模型的对比示意图。

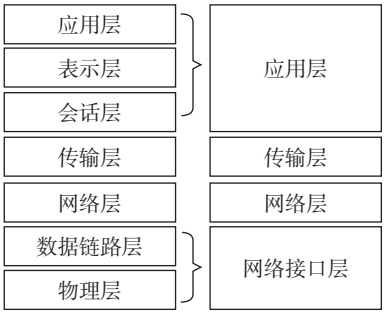


图3.16　OSI和TCP/IP体系结构

OSI 标准制定过程中采用的方法是将整个庞大而复杂的问题划分为若干个容易处理的小问题，这就是分层的体系结构办法。遵照该方法，OSI定义了由下至上的1~7 层，分别为物理层（Physical layer），数据链路层（Data link layer），网络层（Network layer），传输层（Transport layer），会话层（Session layer），表示层（Presentation layer）和应用层（Application layer）。

国际标准化组织ISO制定的OSI参考模型过于庞大、复杂，招致了许多批评。与此对照，由技术人员自己开发的TCP/IP协议栈获得了更为广泛的应用。TCP/IP协议栈是因特网使用的参考模型。在TCP/IP参考模型中，去掉了OSI参考模型中的会话层和表示层（这两层的功能被合并到应用层实现）。同时，将OSI参考模型中的数据链路层和物理层合并为网络接口层。因此，TCP/IP参考模型分为四个层次：应用层、传输层、网络层和网络接口层。

3. 分类

如今，当谈到网络时，通常指如下两种主要类型：局域网（Local Area Network，LAN）和广域网（Wide Area Network，WAN）（见图3.17）。网络归于哪个类型主要取决于其规模。局域网一般是方圆几千米以内的，广域网可以是几十千米甚至是世界范围的。网络规模介于两者之间的，常称为城域网（Metropolitan Area Network，MAN）。随着局域网技术，特别是电信级以太网技术的发展，局域网的应用范围正在向城域网甚至广域网延伸。

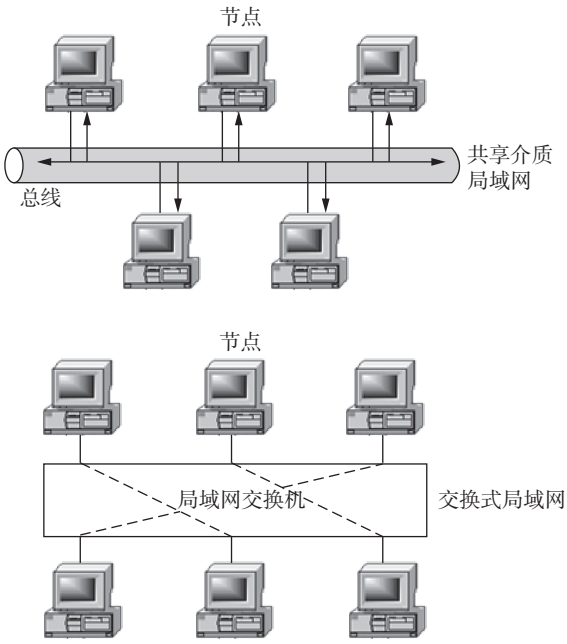


图3.17　共享介质局域网和交换式局域网

（1）局域网

局域网是指在某一区域内由多台计算机互联成的计算机组。局域网通常是专用的，连接的设备往往在同一间办公室、同一建筑物或园区中。局域网可以实现文件管理、应

用软件共享、打印机共享、工作组内的日程安排、电子邮件和传真通信服务等功能。局域网是封闭型的，可以由办公室内的两台计算机组成，也可以由一个公司内的上千台计算机组成。

局域网专用性非常强，具有比较稳定和规范的拓扑结构。常见的局域网拓扑结构有：总线结构、环形结构和星形结构。局域网技术中得到广泛应用的是美国电气和电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）802委员会制定的以太网标准和技术。

早期出现的局域网为共享传输介质的以太网（由集线器连接），对信道的占用采用竞争方式（见图3.17）。随着用户数量的增加，信道冲突加剧、网络性能下降，每个用户实际获得的带宽急剧减小，甚至引起网络阻塞。解决这一问题的方法是在网络中引入双端口或多端口网桥，网桥的作用是将网络划分成多个网段以减小冲突域，提高网络传输性能。由于网桥隔离了冲突域，在一定条件下具有增加网络带宽的作用。但在一个较大的网络中，为保证响应速度，往往要分割很多网段，这不但会增加建设成本，而且将使网络的结构和管理也变得复杂。

目前主流的局域网交换技术是于20世纪90年代初，在多端口网桥的基础上发展起来的。这是一种改进的网桥技术，与传统的网桥相比，它能提供更多的端口，端口之间通过空分交换矩阵或存储转发部件实现互联。网络交换机的引入，既提高了网络性能和数据传输的可靠性，又增强了网络的扩展性。共享介质局域网（传统以太网）与交换式局域网的对比如图3.17所示。

局域网交换机工作在数据链路层，能够读取数据帧中的物理地址，并根据物理地址进行信息交换。由于交换机一般具有高速的交换总线，所以可同时在很多端口之间交换数据。与网桥相比较，局域网交换机具有更多的端口和更高的交换速率。传统网桥大多数基于软件实现，转发时延为毫秒级。局域网交换机的工作大多数由硬件完成，后来通常使用网络处理器实现，转发时延更小，为微秒级。因此，可以实现线速转发，使局域网的交换性能得到明显的改善。

局域网交换机在一定程度上减少了网络冲突，提高了数据转发性能。但由于交换机采用端口地址自动学习和广播相结合的机制没有从根本上改变，仍会导致广播风暴，难以满足大型网络的组网需要，因此又引入了路由器。用路由器来实现不同局域网之间的互联，在不同子网之间转发分组。路由器可以彻底隔离广播风暴，能够适应大型组网对性能、容量和安全性的要求。路由器具有路由选择功能，不但可为跨越不同局域网的分组选择最佳路径，而且可以避开失效的节点或网段，还可以进行不同类型网络协议的转换，实现异构网络的互联。路由器将很多分布在各地的计算机局域网通过广域网互联起来，便可构成互联网，实现更大范围的资源共享和信息传送。目前，最大的计算机互联网就是国际互联网（Internet）。

（2）广域网

广域网又称远程网（Remote Computer Network，RCN），它的作用范围最大，一般可以从几十千米至上万千米。一个国家或国际间建立的网络都是广域网。在广域网内，用于通信的传输装置和传输介质可由电信部门提供。

广域网是由许多交换机组成的，交换机之间采用点到点线路连接，几乎所有的点到点通信方式都可以用来建立广域网，包括租用线路、光纤、微波、卫星信道等。而广域网交换机实际上就是一台计算机，有处理器和输入/输出设备，用于进行数据包的收发处理。

广域网一般最多只包含OSI参考模型的最下面的三层，而且目前大部分广域网都采用存储转发方式进行数据交换，也就是说，广域网是基于报文交换或分组交换技术的网络（传统的公用电话交换网除外）。广域网中的交换机先将发送给它的数据包完整接收下来，然后经过路径选择找出一条输出线路，最后交换机将接收到的数据包发送到该线路上去，以此类推，直到将数据包发送到目的节点。

广域网可以提供面向连接和无连接两种服务模式。对应两种服务模式，广域网有两种组网方式：虚电路（virtual circuit）方式和数据报（datagram）方式，其相关内容见3.4节信息交换技术。

（3）城域网

城域网通常覆盖范围是一个乡镇或城市。它是为需要高速地连接因特网的用户而设计的，端点可分布在整个城市或城市的一部分。其有两个实例：一是可向用户提供高速DSL线路的部分电话网；一是现在可用于因特网的高速数据连接的有线电视网络。

4. 因特网

今天，很少单独看到一个单独的局域网或广域网，它们都是与另一个网络连接的。当两个或多个网络彼此连接，它们就成为互联网络，即互联网（internet）。因特网（Internet）是世界上最大的互联网络（用户数以亿计，互连的网络数以百万计）。因此因特网是由众多广域网、城域网、局域网及单机按照一定的通信协议组成的国际计算机网络。通过因特网，人们可以与远在千里之外的朋友相互发送邮件，共同完成一项工作，或共同娱乐。

因特网是基于TCP/IP协议栈实现的。TCP/IP协议栈由很多协议组成，不同类型的协议又被放在不同的层。其中，位于应用层的协议就有很多，比如FTP、SMTP和HTTP等。只要应用层使用的是HTTP协议，就称为万维网（World Wide Web）。之所以在浏览器里输入某网站的网址时，能看见该网站提供的网页，就是因为网页浏览器和该网站的服务器之间使用的是HTTP协议在交流。

因特网上有着丰富的信息资源。当我们进入因特网后就可以利用其中各个网络和计算机上无穷无尽的资源，与世界各地的人们自由通信和交换信息，以及去做通过计算机能做的各种各样的事情，享受因特网为我们提供的各种服务。因特网提供的主要服务有：电子邮箱、浏览检索、远程访问、大众论坛、信息服务以及多媒体通信等。

3.4　信息交换技术

在由若干网络节点和传输链路按多种拓扑形式连接起来的各种通信网络中，“交换”是一个基本概念，解决的是通信网络中跨节点的数据（这里的数据是指以任何格式表示的信息）传送问题。通信网应为所有进网的数据提供通路，使数据能够顺利地通过网络各中间节点而顺利传送到目的节点。这种自动提供数据通路的技术称为数据交换技术或信息交换技术。目前，根据所交换信息的特征，以及为完成交换功能所采用的不同技术，可分为多种交换技术，包括：电路交换、报文交换、分组交换等。比较新的交换技术包括光交换和软交换。

3.4.1　电路交换

电路交换（Circuit Switching）是最早出现的一种交换技术，主要用于电话业务。电路交换的基本过程包括呼叫建立、信息传送（通话）和连接释放这三个阶段（见图3.18）。

在双方开始通信之前，发起通信的一方（通常称为主叫方）通过一定的方式（如拨号）将被叫方的地址通知网络，网络根据地址在主叫方和被叫方之间建立一条电路，这个过程称为呼叫建立（或称连接建立）。然后主叫方和被叫方可以进行通信（通话），通信过程中双方所占用的通道将不为其他用户使用。完成通信后，主叫方或被叫方通知网络释放通信信道，这个过程

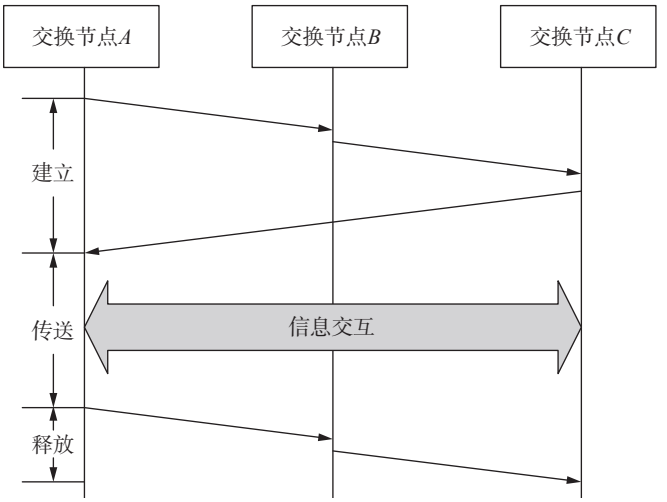


图3.18 电路交换的一般过程

称为呼叫释放（或连接释放）。本次通信过程所占用的相关电路被释放后，即可为其他用户通信所用。这种交换方式就称为电路交换方式。包括最早应用的人工电话在内的电话交换，通常都采用电路交换方式。

由上述过程可以看出，电路交换是一种实时交换，当任一用户呼叫另一用户时，应立即在两个用户之间建立电路连接；如果没有空闲的电路，就不能建立呼叫而遭受损失。因此，对于电路交换而言，应配备足够的连接电路，使呼叫损失率控制在服务质量允许的范围内。

电路交换采用的是固定比特率交换，固定分配带宽（物理信道），在通信前要先建立连接，在通信过程中将一直维持这一物理连接，只要用户不发出释放信号，即使通信（通话）暂时停顿，物理连接仍然保持。即连接建立后，即使没有信息传送也会占用电路，因而电路利用率低。由于通信前要先建立连接，故有一定的连接建立时延；但在连接建立后即可实时传送信息，传输时延一般可忽略不计。电路交换通常采用基于呼叫损失制的方法处理业务流量，超过负荷时呼损率增加，但不影响已建立的呼叫。此外，由于没有差错控制措施，用于数据通信时的可靠性不高。

电路交换的主要优点如下。

（1）信息的传输时延小，对一次接续而言，传输时延固定不变。

（2）信息在通路中“透明”传输，交换机对用户的信息不存储、不分析、不处理，而是原封不动地传送，交换机的处理开销较小，信息的传输效率较高。

电路交换的主要缺点如下。

（1）电路资源被通信双方独占，电路利用率低。

（2）由于存在呼叫建立和连接释放过程，电路的接续时间较长。当通信时间较短（或传送较短信息）时，通信信道建立的时间可能大于通信时间，网络的利用率较低。

（3）有呼损，即可能出现由于被叫方终端设备忙或通信网络负荷过重而呼叫不通的情况。

（4）通信双方在信息传输速率、编码格式、同步方式、通信协议等方面要完全兼容，这就限制了使用各种不同速率、不同代码格式、不同通信协议的用户终端的直接互通。

因此，电路交换技术通常适合于电话交换、文件传送、高速传真等业务，而不适合突发业务和对差错敏感的数据业务。

3.4.2 报文交换

为了克服电路交换技术中各种不同类型和特性的用户终端之间不能互通，电路利用率低及系统有呼损等方面的缺点，提出了报文交换（Message Switching）。20世纪40 年代的电报通信采用了基于存储转发原理的报文交换。

报文交换的基本思想是先将用户的报文存储在交换机的存储器中，当所需要的输出电路空闲时，再将该报文发向接收交换机或用户终端，所以，报文交换又称存储转发交换（Store and Forward），报文交换系统又称“存储–转发”系统。报文交换适合公众电报等。

与电路交换的原理不同，报文交换不需要为通信双方提供实际的物理连接，而是将接收的报文暂时存储，然后按一定的策略将报文转发到目的用户。报文中除了用户要传送的信息以外，还有源地址和目的地址，交换节点要分析目的地址和选择路由，并在该路由上排队，等持有空闲电路时才发送（转发）到下一个交换节点。图3.19示出了报文交换的一般过程。

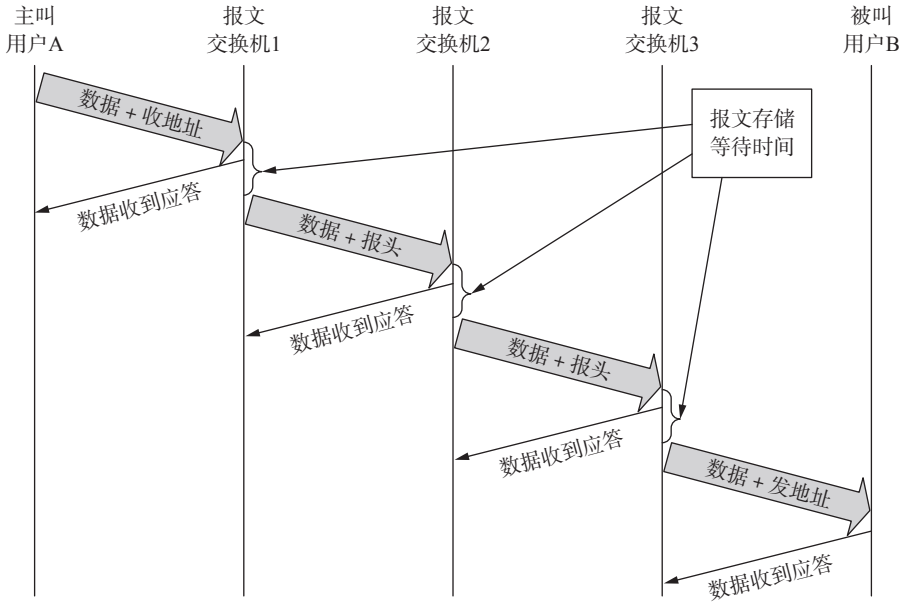


图3.19　报文交换的一般过程

当某一用户A要向另一用户B传送信息（发送报文）时，用户A不需要先叫通与用户B之间的电路，而只需与交换机接通，由交换机暂时把用户A要发送的报文接收和存储起来，并根据报文中提供的用户B的地址确定交换网内的路由，将报文送到输出队列中排队，等到该输出线空闲时立即将该报文发送到下一台交换机，最后根据该报文的路由将报文送达目的用户B。

报文交换中信息的格式以报文为基本单位。一份报文包括三个部分：报头、正文和报尾。报头包括发信站地址（源地址）、终点收信站地址（目的地址）及其他辅助信息；正文为用户要传送的信息；报尾是报文的结束标志，若报文长度有规定，则报尾可以省略。报文交换的特征是交换机要对用户的信息进行存储和处理。

报文交换的主要优点如下。

（1）信息以“存储–转发”方式通过交换机，输入和输出电路的速率、编码格式等可以不同，很容易实现各种不同类型终端之间的相互通信。

（2）在信息传送（报文交换）的过程（从用户A到用户B）中没有电路接续过程，来自不同用户的信息可以在一条线路上以报文为单位进行多路复用，线路可以按它的最高传输能力工作，电路利用率较高。

（3）用户无须叫通目的用户就能发送报文；如果需要，同一报文可以由交换机转发给多个不同的用户，实现多播通信。

报文交换的主要缺点为如下。

(1) 由于“存储–转发”方式，信息通过交换机时产生的时延大，而且时延的变化也较大，不利于实时通信。

(2) 交换机要有能力存储用户发送的报文，其中有的报文可能很长，要求交换机具有高速处理能力和较大的存储容量。

3.4.3 分组交换

电路交换的电路利用率较低，并且不适于不同速率、不同编码方式、不同通信协议用户终端之间的相互通信；报文交换虽然可以进行速率和码型的转换，具有差错控制功能，但在网络负荷层时信息传输时延太长，不适合许多数据通信系统的实时性要求（注意，数据通信系统的实时性要求是指利用计算机进行通信时，用户可以实时地交互传输信息，相比于话音的时延要求，数据通信的实时传输时延要求宽松得多）。分组交换技术较好地解决了这些矛盾。

“分组交换”（Packet Switching）与“报文交换”技术类似，同样采用“存储–转发”方式，但不是以报文为单位，而是把报文划分成许多比较短的、规格化的“分组”（Packet）进行交换和传输。分组长度较短，且具有统一的格式，便于在交换机中存储和处理。分组进入交换机后只在主存储器中停留很短的时间，进行排队处理，一旦确定了新的路由，就很快输出到下一个交换机或用户终端。分组通过交换机或网络的时间很短（为毫秒级），能满足绝大多数数据通信对信息传输的实时性要求。根据交换机对分组的不同处理方式，分组交换有两种工作模式：数据报（Datagram）和虚电路（Virtual Circuit）。

1. 数据报工作模式

数据报方式类似于报文交换，只是将每个分组作为一份报文来对待。每个数据分组中都包含目的地址信息。分组交换机为每一个数据分组独立地寻找路径，因此一份报文包含的多个不同分组可能会沿着不同的路径到达目的地，在目的地需要重新排序。

2. 虚电路工作模式

虚电路方式类似于电路交换。两台用户终端设备在开始传输数据之前，同样必须通过网络建立连接，只是建立的是逻辑上的连接（虚电路），而不是物理连接。一旦这种连接建立之后，用户发送的数据（以分组为单位）将顺序通过该路径传送到目的地。当通信完成之后用户发出拆链请求，网络清除连接。由于分组在网络中是顺序传送的，因而无须在目的地重新排序。虚电路工作模式主要应用于交换式广域网。

综上所述，无论数据报工作方式还是虚电路方式，分组交换的主要优点如下。

(1) 向用户提供了不同速率、不同编码方式、不同通信协议的数据终端之间能够相互通信的灵活的通信环境。

(2) 在网络负荷较轻的情况下，信息的传输时延较小，而且时延的变化不大，能够较好地满足计算机实时交互业务的要求。

(3) 实现了线路的动态统计复用，通信线路（包括中继线和用户线）的利用率很高，在一条物理线路上可以同时提供多条信息通路。

(4) 可靠性高。分组在网络中传输时可以在中继线和用户线上分段进行差错校验，使信息在分组交换网络中传输的误比特率大大降低，一般可以达到10^{-10}以下。由于分组在网络中传输

的路由是可变的，当网络中的设备或线路发生故障时，分组可以自动地避开故障点，故分组交换的可靠性较高。

（5）经济性好。信息以分组为单位在交换机中存储和处理，不要求交换机具有很大的存储容量，降低了网络设备的费用；对线路的动态统计复用也大大降低了用户的通信费用。

分组交换的主要缺点如下。

（1）网络附加的信息较多，对长报文通信的传输效率较低。按照分组交换的要求，一份报文要分割成许多分组在网内传输。为了保证这些分组能够按照正确的路径安全、准确地到达目的地，要给每个数据分组加上控制信息（分组头）。此外，还必须附加许多控制分组，用它们来实现数据通路的建立、保持和拆除，并进行差错控制和流量控制等。可见，在分组交换网内除了传输有用的用户数据之外，还要传输许多辅助控制信息。对于那些长报文而言，分组交换的传输效率可能不如电路交换或报文交换高。

（2）技术实现复杂。分组交换机要对各种类型的分组进行分析处理，为分组的传输提供路由，并且在必要时自动进行路由调整，为用户提供速率、编码和通信协议的转换，为网络的维护管理提供必要的报告信息等，因而技术复杂，要求交换机具有较高的处理能力。

（3）时延较大。由于节点处理任务较多，信息从一端传送到另一端，穿越网络的路径越长、节点越多，分组时延就越大。这种传统的分组交换主要用于数据通信，很难应用于实时多媒体业务。

分组交换是目前应用最广的交换技术。它结合了电路交换和报文交换两者的优点，可使性能达到最优。X.25分组交换、帧中继交换、ATM交换、IP交换、多协议标签交换等都属于分组交换。下面分别进行介绍。

1. X.25分组交换技术

X.25技术用于早期的数据通信网，也就是说，传统的分组交换是基于X.25协议的。X.25协议有三层，第一层为物理层，第二层为数据链路层，第三层为分组层，对应于开放系统互连（Open System Interconnection，OSI）参考模型的下三层，每一层都包含了一组功能。

X.25协议是针对20世纪70年代以模拟通信为主的通信网络环境而设计的。当时可供传输数据的信道大多数是频分制的电话信道，信道带宽为话带带宽（0.3~3.4 kHz），数据传输速率一般不高于9.6 Kb/s，误比特率（误码率）为10^{-4}~10^{-5}。这样的误码率无法满足数据通信的要求。在这种环境下设计的数据传输协议X.25，为了兼顾网络的效率和传输的可靠性，采用了诸如差错控制、停（止）等（待）重发、流量控制等措施，因而协议比较复杂。

2. 帧中继交换技术

20世纪80年代以后，随着光通信技术的发展，光纤逐渐成为通信网传输介质的主流。光纤通信具有容量大、质量高的特点，数据传输误码率小于10^{-9}，系统能够提供10~100 Gbit/s的数据传输速率。在这样的通信环境下运行分组协议，显然没有必要像X.25协议那样再做许多精巧而烦琐的控制。快速分组交换（Fast Packet Switching，FPS）就是在这样的背景下提出的。快速分组交换可理解为尽量简化协议，只保留核心的链路层功能，以提供高速、高吞吐量、低时延服务的交换方式。

帧中继就是一种在数据链路层使用简化的方式传送和交换数据单元的协议。它简化了X.25协议的差错检验、数据流控制等功能，只有下两层，没有第三层，在数据链路层也只保留了核心功能，如帧的定界、同步及差错检测等。与传统的分组交换相比，帧中继有如下几个主要特点。

（1）帧中继以帧为单位来传送和交换数据，在第二层（数据链路层）进行复用和传送，而不是在分组层，这样就简化了协议，加快了处理速度；

（2）帧中继将用户面与控制面分离，而通常的分组交换是不分离的，用户面负责用户信息的传送，控制面负责提供呼叫控制和连接管理，包括信令功能。

（3）帧中继取消了X.25协议中规定的网络节点之间、网络节点与用户设备之间每段传输链路上的数据差错控制，将逐段链路上的差错控制推到了网络的边缘，由终端负责完成。网络只进行差错检测，不进行差错控制，减少了处理环节，提升了速度。帧中继的这种处理思路非常适应数据传输的误码率很低的光纤通信。

3. ATM交换技术

从前面的介绍中可以看出，电路交换和分组交换具有各自的优势和缺陷，两者实际上是互补的。电路交换适合实时业务，但是无法适应不同速率业务的要求，并且网络的利用率较低；而分组交换可以适配各种速率业务，具有较高的网络利用率，却无法很好地支持实时业务。显然，能够适应各种不同业务要求，并能够支持多媒体通信的交换和复用技术必须综合电路交换和分组交换的优势，支持高速和低速的实时业务，具有高效的网络利用率，这正是国际电信联盟电信标准化部门（International Telecommunications Union Telecommunication Standardization Sector，ITU-T）提出异步传输模式（Asynchronous Transfer Mode，ATM）的初衷。ATM也是一种快速分组交换技术，其基本特点如下。

（1）采用固定长度的信元

与采用可变长度帧的帧中继比较，ATM交换采用固定长度的信元（Cell）作为交换和复用的基本单元。信元实际上就是长度很短的分组，只有53字节（Byte），其中前5字节称为信头（Cell Header），其余48字节为信息域，或称为有效载荷（Payload）。采用长度很短的信元可以减小交换节点内部的缓冲器容量，以及排队时延和时延抖动。信元的长度固定，则有利于简化交换控制和缓冲器管理。

信头中包含控制信息的多少反映了交换节点的处理开销。因此，要尽量使信头简化，以减少处理开销。ATM信元的信头只有5字节，主要包括虚连接的标志、优先级标志和信头的差错校验等。信头中的差错校验是针对信头本身的，这是非常必要的功能，因为信头如果出错，将导致信元丢弃或错误选路。

（2）面向连接

ATM采用的是分组交换中的虚电路方式，即采用面向连接的方式。在用户传送信息之前，先要有连接建立的过程；在信息传送结束之后，要断开连接。这一点与电路交换方式类似。当然，这里不是物理连接，而是一种虚连接。

为了便于应用和管理，ATM的虚连接分成两个等级：虚信道（Virtual Channel，VC）和虚通路（Virtual Path，VP）。传输通道可包含若干个虚通路，每个虚通路又可划分为若干个虚信道（见图3.20）。

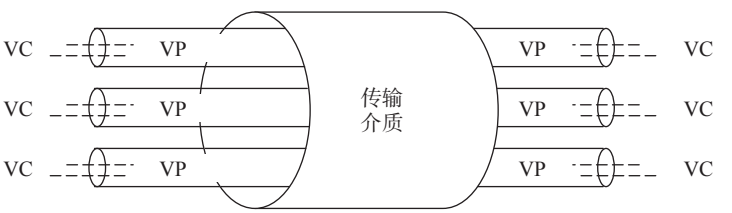


图3.20　传输通道示意图

（3）异步时分交换

采用电路交换方式的数字程控电话交换属于同步时分（Synchronous Time Division，STD）交换，ATM交换则属于异步时分（Asynchronous Time Division，ATD）交换。

为了说明ATD的概念，先介绍STD的概念。时分意味着复用，即一条物理链路可以由多个连接所共享，各自占用不同的时间位置。各个连接属于不同的呼叫，有各自的目的地，在交换的过程中必须加以区分，也就是说，要判别每个时间位置中的信息是属于哪个连接的。STD的特点是，按照时间位置本身来区分，这意味着每个连接占有某条物理链路上固定的时间位置。以脉冲编码调制（Pulse Code Modulation, PCM）体制的E1标准为例，每帧有32个时隙，假如在呼叫建立过程中将第10个时隙分配给连接A，则每帧的第10个时隙始终是传送连接A的用户信息，周而复始，直到连接拆除。

ATD复用的各个时间位置相当于各个信元所占的位置，即一个信元占有一个时间位置。ATD与STD不同的是，属于某个呼叫连接的多个信元不占有固定的时间位置，而是根据该呼叫连接所需的带宽大小，占有或多或少的时间位置。也就是说，属于同一呼叫连接的信元可以或密或疏地在复用链路上出现。因此，它不是固定分配时隙的同步方式，而是灵活分配带宽的异步方式，因而可以适应各种不同带宽业务的需求。

为便于比较，图3.21简明地示意了STD与ATD概念的区别。

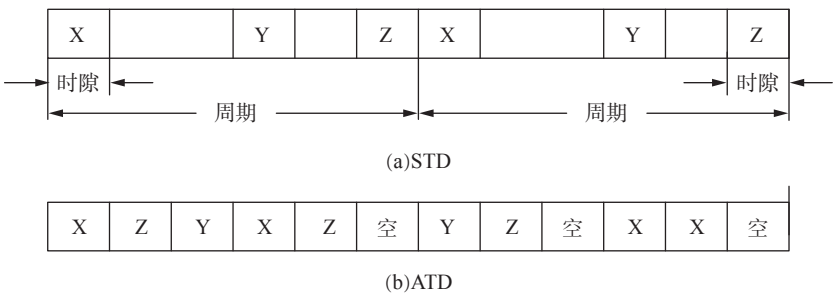


图3.21　STD与ATD的概念

图3.21(a)所示为STD，X，Y和Z表示不同的呼叫连接，它们占用了各自的固定时隙位置；图3.21(b)所示为ATD，X，Y和Z表示不同虚连接所属的信元，它们的时间位置是灵活分配的。

从上面的描述可以知道，实际上ATM交换充分地综合了电路交换和分组交换的优点，它既具有电路交换方式“处理简单”的特点，支持实时业务，数据透明传输（网络内部不对数据进行复杂处理）并采用端到端的通信协议，同时也具有分组交换方式支持可变比特率业务的特点，并能对链路上传输的业务进行统计复用。因而，ATM很快为人们所接受，成为早期宽带网络交换和复用的首选技术。

从技术角度看，ATM在多业务承载方面是最佳的，而且ATM相关协议和标准十分完善，但其协议体系的复杂性造成了ATM系统的研制、配置、管理、故障定位都具有难度。在当时情况下，ATM没有机会将原有设施推倒重来，构建一个纯ATM网。相反，ATM必须支持已经应用到桌面的网际协议（Internet Protocol，IP）才能够生存。传统的IP技术只能提供尽力而为（Best Effort）服务，没有任何有效的服务质量保证机制。同时，IP技术在发展过程中也遇到了路由器瓶颈等问题。如果把ATM与IP技术结合起来，既可以利用ATM网络资源为IP用户提供高速数据转发，发展ATM上的IP用户业务，又可以解决Internet发展中的瓶颈问题，推动Internet业务的进一步发展。ATM的发展在20世纪90年代中期达到顶峰。也就是在此期间，世界通信技术及网

络技术的发展格局逐渐发生了重大变化，特别是Internet的发展，使ATM的应用受到很大影响。ATM缺乏业务和末端用户支持、价格昂贵、技术复杂的缺点日益显现，IBM公司力图使ATM技术走向桌面的努力也未能取得成功，最后，ATM技术的发展与应用被IP的简单、灵活和经济性所淹没。

由于技术复杂和成本方面的原因，虽然ATM在商业用途中没有得到大范围应用，但是其具有支持综合业务及确保服务质量（Quality of Service，QoS）的特点，使得它在军事应用中具有不可替代的作用。ATM支持从接入到核心骨干的全系列网络解决方案，非常适合军事应用。在一个指挥所或阵地，一台ATM交换机即可完成数据、话音、图像、视频等各种军事信息的接入和传输，有效地减少了通信设备的种类和数量，适合部队的机动部署和展开。目前世界上许多军队的现役装备都采用了ATM技术。

4. IP交换技术

以Internet为代表的IP技术的迅猛发展，迫切需要提高IP网络的服务质量。传统IP路由器和X.25分组交换机都是在第3层进行转发的，采用软件控制将分组从一个端口转移到另外一个端口，这是基于存储–转发概念的，转发时延较大、速率较低。为了提高IP分组转发的速度，适应数据及多媒体业务发展的需要，IP交换技术应运而生。

IP交换的概念，最早由美国Ipsilon公司在1996年提出。它将IP路由处理器捆绑在ATM交换机上，去除了交换机中原有的ATM信令。IP交换机使用IP路由协议进行路由选择。它的连接建立是由数据流驱动的，即“一次路由、多次交换”：对于单个的IP分组，采用传统IP逐跳转发方式进行转发；对于长持续时间的业务流，能自动建立一个虚通路，使用ATM交换方式进行转发。

综上所述，IP交换技术是Internet采用的技术体制，能很好地支持数据业务，但在保证业务QoS上有一定的困难。它的基本特点是：不需要建立连接、路由不受控制、只能尽力而为地保证QoS、资源利用率高，属于有冲突的传递。

5. 多协议标签交换技术

在IP交换的发展过程中，互联网工程任务组（Internet Engineering Task Force，IETF）起到了积极的推动作用，IETF在1997年初成立了多协议标签交换（Multiple Protocol Label Switch，MPLS）工作组，综合了Cisco和Ipsilon等的IP交换方案，制定出了一个统一的、完善的第3层IP交换技术标准，即MPLS。MPLS明确规定了一整套协议和操作过程，最终通过ATM、帧中继、点对点协议（Point to Point Protocol，PPP）和以太网等实现IP网络快速交换。

MPLS采用了ATM的高效传输交换方式，抛弃了复杂的ATM信令，无缝地将IP技术的优点融合到ATM的高效硬件转发中；在MPLS域中，数据传输和路由计算是分开的，是一种面向连接的传输技术，能够提供有效的QoS保证；MPLS不但支持多种网络技术，而且还是一种与链路层无关的技术，它同时支持X.25、帧中继、ATM、SDH、DWDM等技术，保证了多种网络的互通，可以使不同的网络传输技术统一在一个MPLS上。

综上所述，MPLS所具有的面向连接、高速交换、支持QoS、扩展性好等特点，使它在国内外具体组网中获得了广泛的应用，已成为主流的宽带交换技术。

3.4.4　交换新技术

1. 光交换技术

随着光通信技术的不断进步，波分复用系统在一根光纤中已经能够提供几百吉比特每秒到太比特每秒的信息传输能力。传输系统容量的快速增长给交换系统的发展带来了巨大的压力和动力。通信网交换系统的规模越来越大，运行速率也越来越高，未来的大型交换系统将需要处理总量达几百、上千太比特每秒的信息。但是目前的电子交换和信息处理网络的发展已接近了电子器件的极限，其中所固有的钟偏、漂移、串话、响应速度慢等缺点，都限制了交换速率的进一步提高。为了解决电子器件的瓶颈问题，通信网开始在交换系统中引入光子技术，实现光交换。

光交换技术能够在光域直接将输入的光信号交换到不同的输出端，完成光信号的交换。光交换的优点在于，光信号在通过光交换单元时，无须经过光电、电光转换，因此它不受检测器、调制器等光电器件响应速度的限制，对比特速率和调制方式透明，可以大大提高交换单元的吞吐量。目前，光网络已经由过去的点到点波分复用发展到面向连接的光分插复用（Optical Add-Drop Multiplexer，OADM）和自动交换光网络（Automatically Switched Optical Network，ASON），并将向融合电路交换和分组交换的智能光网络演进。光交换将是未来宽带网络使用的另一种宽带交换技术。

2. 软交换技术

软交换的概念最早起源于美国。当时在企业网络环境下，用户采用基于以太网的电话，通过一套基于PC服务器的呼叫控制软件（CallManager、CallServer），实现程控交换机PBX功能（即IP PBX）。对于这样一套设备，系统无须单独铺设网络，只通过与局域网共享即可实现管理与维护的统一，综合成本远低于传统的PBX。由于企业网环境对设备的可靠性、计费和管理要求不高，主要用于满足通信需求，设备门槛低，许多设备商都可提供此类解决方案，因此IP PBX应用获得了巨大成功。受到IP PBX成功的启发，为了提高网络综合运营效益，使网络的发展更加趋于合理、开放，从而更好地服务于用户，业界提出了这样一种思想：将传统的交换设备部件化，分为呼叫控制与媒体处理，两者之间采用标准协议（MGCP、H248）且主要使用纯软件进行处理，于是，SoftSwitch（软交换）技术应运而生。

软交换技术是一个分布式的软件系统，可以在基于各种不同技术、协议和设备的网络之间提供无缝的互操作性。其基本设计原理是设法创建一个具有很好的伸缩性、接口标准性、业务开放性等特点的分布式软件系统，独立于特定的底层硬件/操作系统，并能够很好地处理各种业务所需的同步通信协议，在一个理想的位置上把该架构推向摩尔曲线轨道。

软交换的实现目标是在媒体设备和媒体网关的配合下，通过计算机软件编程的方式，实现对各种媒体流进行协议转换，并基于分组网络（IP/ATM）的架构实现IP网、ATM网、PSTN网等的互连，以提供与电路交换机具有相同功能并便于业务增值和灵活伸缩的设备。

软交换技术区别于其他技术的最显著特征，也是其核心思想的三个基本要素如下。

（1）开放的业务生成接口

软交换提供业务的主要方式是通过API与“应用服务器”配合，以提供新的综合网络业务。与此同时，为了更好地兼顾现有通信网络，它还能够提供智能业务。

（2）综合的设备接入能力

软交换可以支持众多的协议，以便对各种各样的接入设备进行控制，最大限度地保护用户投资并充分发挥现有通信网络的作用。

（3）基于策略的运行支持系统

软交换采用了一种与传统OAM（Operation Administration and Maintenance，操作、管理与维护）系统完全不同的基于策略（Policy-based）的实现方式来完成运行支持系统的功能，按照一定的策略对网络特性进行实时、智能、集中式的调整和干预，以保证整个系统的稳定性和可靠性。

作为分组交换网络与传统PSTN网络融合的全新解决方案，软交换将PSTN的可靠性和数据网的灵活性很好地结合起来，是新兴运营商进入话音市场的新的技术手段，也是传统话音网络向分组话音演进的方式。目前在国际上，软交换作为下一代网络（NGN）的核心组件，已经被越来越多的运营商所接受和采用。

思考题

1. 通信系统的一般模型是由哪几部分组成的？各部分的作用是什么？
2. 什么是码元传输速率？什么是信息传输速率？两者之间有什么区别？
3. 光纤导光的原理是什么？光纤通信有什么优点？
4. 微波通信系统是如何组成的？微波通信有什么特点？
5. 卫星通信系统是如何组成的？卫星通信有什么优点？
6. 信息网络的类型主要有哪些？信息网络主要由哪些要素构成？
7. 分组交换的原理是什么？
8. 分组交换技术有几种？分别有什么特点？

第4章　信息存储与管理技术

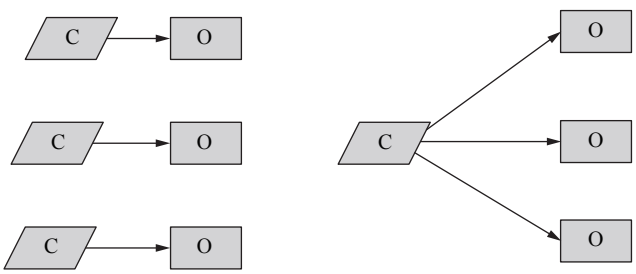
大部分信息需要经过处理才能被再次利用，目前主要的处理方式是依托计算机来实现。因此，为了保证信息能被计算机正确识别，必须要先将其数字化，即对信息进行分类与编码，最终使所有信息具备统一的数字化形式，同时采用相应的存储技术将这些信息存储到计算机中。

本章主要从信息的逻辑存储与物理存储两个方面进行相关技术的介绍，其中逻辑存储部分主要介绍数据库技术与数据仓库技术；物理存储部分主要介绍磁盘阵列技术与网络存储技术。

4.1　信息编码

提到编码，人们就自然而然地想到由数字、字母、连接符等符号组成的符号串，它们与事物对象或事物分类的类目对应，是事物或事物类的代表。本书仅讨论符号形式的编码，并将用符号进行编码所得到的符号串称为代码。

代码与事物对象的关系分为一对一和一对多的关系，如图4.1所示，其中C代表代码（code），O代表事物对象（object）。



(a)代码与事物对象一对一　　(b)代码与事物对象一对多

图4.1　代码与事物对象的关系示例

当代码与事物对象存在一对一的关系时，代码就唯一地代表一个事物对象，这样的代码称为标识码。例如，大学生一进入学校，就有一个唯一标识号，即学号。中华人民共和国的每一个合法公民都有标识号，即身份证号。

当一个代码对应多个事物对象时，可以认为该代码对应于一个事物集合。这个集合并不是由若干个事物对象随便凑在一起的，如果是这样的话，也就没有必要为这个集合编码了。在实际应用中，这个集合是由具有相同或相似特征的事物组成的，编码是针对事物特征的，所以称它为特征码。对特征进行编码有下面几种常见的情况（见图4.2）。

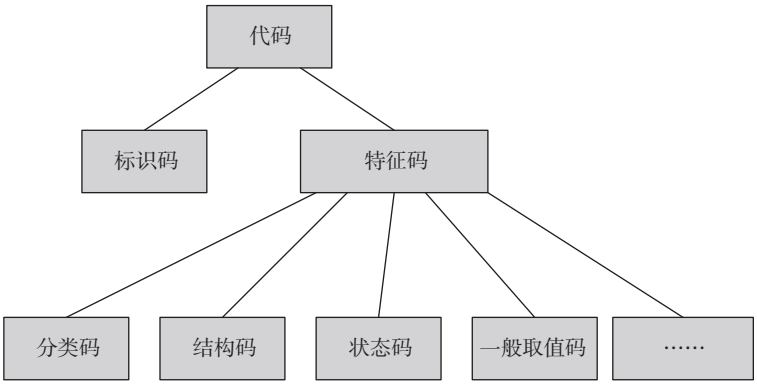


图4.2 代码按其与对象的映射关系和作用划分而成的种类

1. 分类码

代码所表示的集合是由一类事物组成的，代码对应于一个类目。将代表类目的代码或代表一类事物的代码称为分类码。

2. 结构码

结构码用数字来表示事物对象之间的结构关系，表示一个事物对象或一类事物对象在结构中的位置。由于树形结构容易编码和使用，所以一般结构码指的是树形结构码，即每个代码代表事物对应节点在树形结构中所处的位置。

3. 状态码

在对事物进行管理时，常常需要用编码的形式来记录事物所处的状态，这类编码称为状态码。

4. 一般取值码

特征的取值往往限制在可枚举的范围内，有些取值不代表分类，也不说明结构关系，也不算是状态，完全是为了便于计算机表达事物的特征，处理信息。比如在第2章中涉及的侦察信息，有的是以文字形式表示的，有的是以图像形式表示的，还有的是以音频形式表示的，所有的这些信息如果最终都要被计算机接受并处理，就都要进行数字化处理，也就是要实现对信息的“一般取值码”的设定。设定的最终目的是便于计算机对其进行处理。

4.2 数据库技术

计算机发明的最初目的是进行科学计算，所涉及的主要是类型单一但计算复杂的数据，随着计算机技术的发展和应用的深入，其作用已由单纯的数学计算转向复杂的数据处理，可实现从数据中提炼信息，为人们的行为提供决策依据。

在大多数情况下，计算机信息处理的对象是现实生活中的客观事物。在对客观事物实施处理的过程中，首先要经历了解、熟悉的过程，通过观测抽象出大量描述客观事物的信息，再对这些信息进行整理、分类和规范，进而将规范化的信息数据化，最终由数据库系统存储和处理。

在数据处理中，数据描述将涉及不同的范畴。从事物的特性到计算机中的具体表示，涉及了三个层次，经历了两次抽象和转换（见图4.3）。

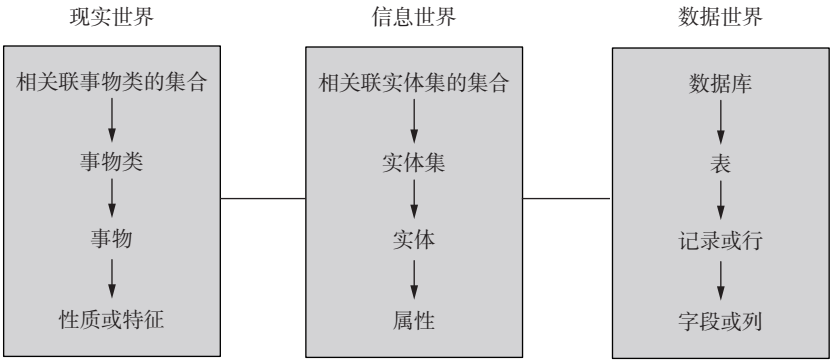


图4.3　数据描述的三个层次

为了让计算机能够处理现实世界中各种客观事物及其之间的相互联系，数据库设计人员首先需要将客观事物抽象为信息世界的概念模型，然后再进一步抽象为计算机中的数据模型。概念模型将客观事物描述为一种信息结构，这种信息结构与具体的计算机结构无关。数据模型是在概念模型的基础上按计算机的表达逻辑和处理能力对数据建模，也就是转换为计算机能理解的概念，如字段、记录、文件、关键字等，它们有严格的形式化定义，以便在数据库中管理和处理这些数据。

4.2.1　信息世界

信息世界就是现实世界在人们头脑中的反映，又称为观念世界。客观事物在信息世界中称为实体，反映事物之间关系的是实体模型或概念模型。现实世界是物质的，反之，信息世界是抽象的。

在信息世界中，通常涉及以下基本概念。

1. 实体

现实世界中的客观事物在信息世界中称为实体。实体既可以指具体的人、事、物，例如一位医生、一名病人、一间教室，也可以是抽象的概念或联系，例如一门课、销售人员与所在部门的工作关系。

2. 属性

实体所具有的某一特性称为属性。一个实体可以用若干个属性来描述。例如，医生实体可以由身份证号、姓名、性别、出生年月、所在科室等属性组成，即（130103198101230868，张山，男，1981/1，心内科）这些属性组合起来表征了一位医生。

属性有型和值之分，属性的“型”就是属性名及其取值类型，属性的“值”就是属性在其值域中所取的具体值。例如，医生实体中的“姓名”属性，“姓名”和取值“字符类型”就是其“型”，而“张山”则是其“值”。

3. 实体标识符

能够唯一标识一个实体的属性集称为实体标识符，也称为键或码，如身份证号是医生实体的键。

4. 域

属性的取值范围称为该属性的域，例如身份证号的域为18位整数，姓名的域为字符串集合，性别的域为{男，女}等。

5. 实体型

具有相同属性的实体必然具有共同的特征和性质。实体名及其属性名集合用来抽象和刻画同类实体，即实体的结构描述，称为实体型。例如，医生（身份证号，姓名，性别，出生年月，所在科室）就是一个实体型。

6. 实体集

同型实体的集合称为实体集。例如，某个医院的全体医生就是一个实体集。

7. 联系

在现实世界中，事物是相互联系的，这种联系必然要在数据库中有所反映。建立概念模型的一个主要任务就是确定实体之间的联系。所谓的联系，既包括实体内部的联系，也包括实体间的联系。实体内部的联系通常指的是组成实体的各属性之间的联系，实体间的联系通常是指不同实体集之间的联系。

4.2.2 数据世界

数据世界就是信息世界中的信息数据化后对应的产物。现实世界中的客观事物及其联系，在数据世界中是以数据模型来描述的。前文提到了信息世界中的实体可抽象为数据世界中的数据，而这些数据将存储在计算机中。在数据世界中常用到以下一些基本概念。

1. 字段

信息世界中的“属性”在数据世界中相应地称为字段，也称为数据项。字段的命名往往和属性名相同。例如，医生会有“身份证号、姓名、性别、出生年月、所在科室”等字段。

2. 记录

信息世界中的每个“实体”相应地均可以数据化为每条记录，如一名医生（130103198101230868，张山，男，1981/1，xn001）为一条记录。

3. 表

表对应信息世界中的“实体集”，在数据世界中指的是所有实体数据化后的所有记录的集合。如所有医生的记录就组成了一个医生表。

总之，客观事物是信息之源，是设计、建立数据库的出发点。而概念模型和逻辑模型是对客观事物及其相互关系的两种抽象描述，实现了信息处理三个层次的对应转换。也就是说，现实世界到信息世界的第一层抽象是依靠概念模型进行建模分析，而信息世界到数据世界的第二层抽象是依靠逻辑模型进行建模分析。在下述小节中将依次介绍概念模型和逻辑模型的相关知识。

4.2.3 概念模型

数据模型是数据库系统的核心和基础。为了把现实世界中的具体事物抽象、组织为某一数据库管理系统支持的数据模型，人们首先将现实世界中的客观对象抽象为某一种信息结构，这

种信息结构并不依赖于具体的计算机系统，也不是某一个数据库管理系统支持的数据模型，而是概念层面的模型，然后再把概念模型转换为计算机上某一数据库管理系统支持的数据模型，这一过程如图4.4所示。由此可见，概念模型主要用于信息世界的建模，是现实世界到信息世界的第一层抽象，是数据库设计人员进行数据库设计的有力工具，也是数据库设计人员和用户之间进行交流的语言。

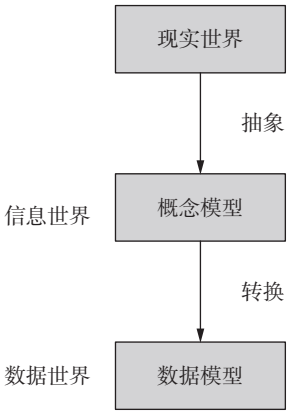


图4.4　从现实世界到数据世界的转换过程

4.2.3.1　概念模型的基本概念

概念模型是对信息世界的建模，因此它应该能够方便、准确地表示出信息世界中的常用概念。表示概念模型的方法很多，其中最著名和常用的是P. P. S. Chen于1976年提出的E-R（实体–联系）方法，该方法用E-R图来描述概念模型，E-R方法也称为E-R模型。

在E-R模型中采用了三个主要概念，即实体、属性和实体联系。

1. 实体

现实世界中客观存在并可相互区别的“事物”称为实体。每个实体有一组性质，其中一部分性质的取值可以唯一地标识实体。例如，中国的每个公民都会有一个身份证号，唯一地标识这个公民。实体可以是实实在在的，如人或物体；也可以是抽象的，如成绩、课程或概念。

2. 属性

实体通过一组属性来表示，属性是实体集中每个成员具有的描述性性质。一个实体往往有多个属性，这些属性之间是有关系的，它们构成该实体的属性集合。如果一个属性或几个属性的子集合能够唯一标识整个属性集合，则称该属性子集为属性集合的标识码或键。实体的属性集合可能会存在多个键，每一个键都可以称为候选键，但一个属性集只能定义一个唯一的标识码。一旦某个候选键被选为唯一标识，就称其为该属性集的主键（或主码）。如果一个实体的某个属性集合本身不是该实体的键，而是另一实体的键，则称其为外键。外键描述了两个实体之间的联系。

例如，实体集doctor可能具有属性dno、dname、dgender、dbirthday、dunit和doffice；每个实体的所有属性都有一个值。例如，对于某个特定的doctor实体，它的dno值为2013h00101，dname值为张山，dgender值为男，dbirthday值为1981-1-1，dunit值为h001，doffice值为00006。

dno属性用来唯一标识医生，因为可能存在具有相同名字、性别、出生日期和科室的医生，因此dno及包含dno的集合都可称为候选键，最终确定dno作为唯一标识，则dno为该属性集的主键。

dunit本身并不是doctor实体的键，但却是另一实体unit（科室）的键，则称dunit为doctor实体的外键，它表示了doctor实体与unit实体之间的联系。如表4.1显示了医院管理数据库中的实体集doctor。

表4.1　实体集doctor

2013h00101	张山	男	1981-1-1	h001	00006
2013h00102	李梅	女	1982-1-1	h001	00007
2013h00103	王武	男	1982-2-2	h001	00008
……	……	……	……	……	……

当实体在某个属性上没有值时使用Null（空）值，Null值可以表示“不可用”，即该实体的这个属性值不存在，也可以表示属性未知。

3. 实体联系

常见的实体联系有三种（见图4.5）。

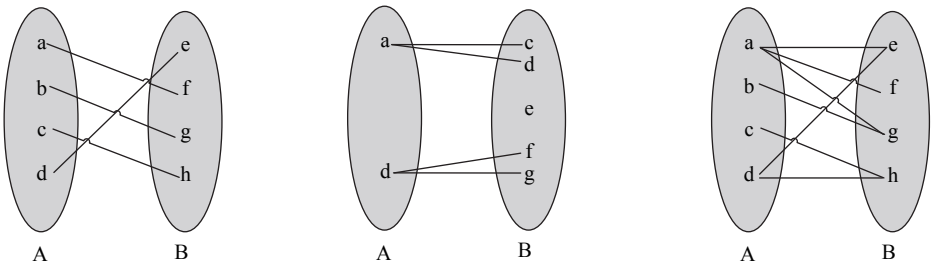


图4.5 不同实体的集实体之间的联系

（1）一对一联系（1∶1）

如果实体集A中的一个实体至多与实体集B中的一个实体相对应（相联系），反之亦然，则称实体集A与实体集B的联系为一对一的联系。如果一个班级只能有一个班长，一个班长也只能在一个班级任职，则班级与班长之间是一对一的联系。

（2）一对多联系（1∶n）

如果实体集A中的一个实体与实体集B中的多个实体相对应，反之，实体集B中的一个实体至多与实体集A中的一个实体相对应，则称实体集A与实体集B的联系为一对多的联系。例如，一个母亲可以有多个子女，而一个子女只会有一个母亲，母亲与子女的联系即为一对多的联系。同理，学校对系、系对班级、班级对学生都是一对多的联系。

（3）多对多联系（m∶n）

如果实体集A中的一个实体与实体集B中的多个实体相对应，而实体集B中的一个实体也与实体集A中的多个实体相对应，则称实体集A与实体集B的联系为多对多的联系。例如，一个老师可以有多个学生，而一个学生同时也会有多个老师，老师与学生的联系即为多对多的联系。

4.2.3.2 概念模型的表示方法

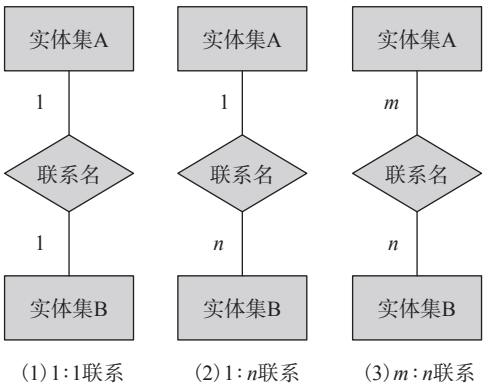
在概念模型中，通常用E-R图表示实体型、属性和联系。具体的表示方法如下。

实体型：用矩形表示，矩形框内写明实体名。

属性：用椭圆形表示，并用无向边分别与有关实体连接起来。

联系：用菱形表示，菱形框内写明联系名，并用无向边分别与有关实体连接起来，同时在无向边旁标上联系的类型（1∶1，1∶n，m∶n）。需要注意的是，如果一个联系具有属性，则这些属性也要用无向边与这个联系连接起来。

用于表示两个实体集之间的三种联系的E-R模型如图4.6所示。



（1）1∶1联系　（2）1∶n联系　（3）m∶n联系

图4.6 实体集之间的三种联系的E-R模型

例如在校园系统中，学校实体与校长实体之间存在1∶1的联系，其E-R模型图如图4.7所示。

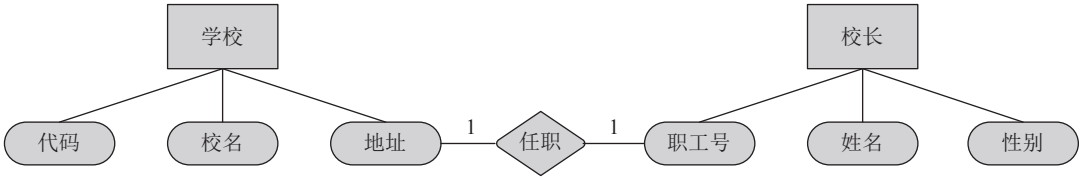


图4.7　1∶1的E-R模型图

学校招收若干学生，产生了1∶*m*的联系，其E-R图如图4.8所示。同时，在表示联系时，除了标明联系的名称之外，还要表示联系的类型。图中可以用任何字母表示“多”。在该例中，当学校招收学生时就会产生学生的入校时间等一些新的属性（在图中仅标出入校时间这一个属性作为代表）。而这样的属性只有在学校实体和学生实体产生联系时才出现，它们既不是学校这个实体的属性，也不能归于学生的个人信息的属性之中，因此把这样的属性加在“招收”这个联系上。

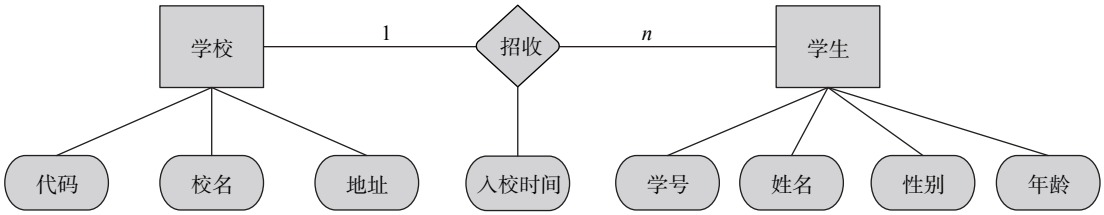


图4.8　1∶*m*的E-R模型图

类似地，学生和课程之间存在*m*∶*n*的联系，即多对多的关系，一名学生可以选择多门课程，每门课程可以由多个学生选择学习，其E-R模型图如图4.9所示。

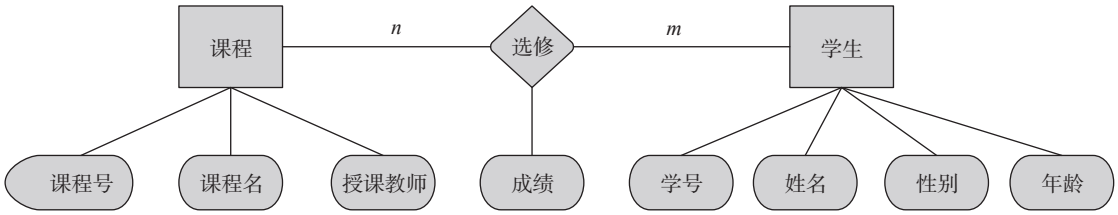


图4.9　*m*∶*n*的E-R模型图

同上例一样，学生选修课程后产生了一个新的属性：每个学生对应着每门课程的成绩，应该把这样的属性附加在“选修”这个联系上。

4.2.4　逻辑模型

在建立了概念模型后，接下来就可将建好的概念模型转换为逻辑模型。逻辑模型是用户在数据库中看到的模型，是具体的数据库管理系统所支持的数据模型，主要用于数据库管理系统的实现。该模型反映了系统分析人员对数据存储的观点，是对概念模型的进一步分解和细化。在逻辑模型中，最常用的是关系模型，20世纪80年代以来，计算机厂商新推出的数据库管理系统几乎都支持该模型。

关系模型主要由数据结构、数据操作和数据完整性约束这三个要素组成。其中，数据结构描述数据对象的类型、内容、性质以及数据间的联系等静态特征，同时也决定了相应的数据组织方式；数据操作定义了在数据库中执行的数据操作的含义、符号、操作规则及实现操作的语

言；完整性约束定义了对给定数据模型中的数据及联系的一组制约和依存规则，用以保证数据库中数据的正确、有效、相容。

4.2.4.1 关系模型的数据结构

1. 域和笛卡儿积

在介绍关系之前，先了解一下“域”和“笛卡儿积”这两个概念。其中，域是一组具有相同数据类型的值的集合，例如，整数、实数、指定长度的字符串集合、介于某个范围的日期等。

给定一组域$D_1, D_2, \cdots, D_n$，其中可以有相同的域，则$D_1, D_2, \cdots, D_n$的笛卡儿积为：

$$D_1 \times D_2 \times D_3 \times \cdots \times D_n = \{(d_1, d_2, \cdots, d_n) | d_i \in D_i\}$$

假定有两个域D_1和D_2，分别为教师集合和课程集合，其中每名教师均具备讲授每门课程的能力：

D_1 = {潘越，徐华，倪涛}，D_2 = {数学，语文，英语}，则其笛卡儿积为

$D_1 \times D_2$ = {（潘越，数学），（潘越，语文），（潘越，英语），
（徐华，数学），（徐华，语文），（徐华，英语），
（倪涛，数学），（倪涛，语文），（倪涛，英语）}。

该笛卡儿积的基数为$3 \times 3 = 9$，即$D_1 \times D_2$一共有9个元组。这9个元组可列成一个二维表，它代表教师和课程之间所有可能的组合，如表4.2所示。

表4.2 两个关系的笛卡儿积

教师	课程
潘越	数学
潘越	语文
潘越	英语
徐华	数学
徐华	语文
徐华	英语
倪涛	数学
倪涛	语文
倪涛	英语

实际上，两个域的笛卡儿积所产生的9个元组在现实中不一定都是有意义的，所以一般来说，只有笛卡儿积子集能反映现实世界，具有实际意义。

在以上理解的基础上，接下来可以讨论关系的概念。

$D_1 \times D_2 \times D_3 \times \cdots \times D_n$的子集称为在域$D_1, D_2, \cdots, D_n$上的关系，表示为：

$$R(D_1, D_2, \cdots, D_n),$$

其中R表示关系名，n表示关系的目或度。关系的成员为元组，即笛卡儿积的子集元素$(d_1, d_2, \cdots, d_n)$，其中d_i为元组的第i个分量。既然关系为笛卡儿积的有限子集，则关系也是一个二维表，表的每行对应一个元组，每列对应一个域。为了区分每个域，则对每列起一个名字，称为属性。

上述例子中的笛卡儿积$D_1 \times D_2$表示教师和课程之间所有可能的组合，而在某个学期里这些教师正在讲授的课程则是$D_1 \times D_2$的一个子集。例如：

$$R = \{(潘越, 数学), (徐华, 语文), (倪涛, 英语)\}$$

这就是一个关系，代表教师正在讲授的课程。

2. 关系模型

关系模型的数据结构单一，是建立在严格的数学概念基础上的。在普通用户看来，关系模型中的数据结构是一个二维表，由行和列组成。现以学生信息登记表为例，介绍关系模型中的一些术语。

表4.3　学生信息登记表

姓名	学号	民族	出生日期	党/团时间	籍贯	班次	职务	固定电话	家庭住址
潘越	3912006001	汉	1987.10.21	2010.6.7	湖南省	5班	学员	3513934	……
徐华	3912006002	汉	1987.10.21	2010.6.7	福建省	1班	学员	3513934	……
倪涛	3912006003	汉	1987.10.21	2010.6.7	四川省	3班	副班长	3513934	……
……									

（1）关系。一个关系通常对应一个表，如以上所示学生信息登记表。

（2）元组。表中的一行即为一个元组。

（3）属性。表中的一列即为一个属性，给每个属性起一个名称即为属性名，例如，在上表中有10个列，对应10个属性（姓名，学号，民族，出生日期，党/团时间，籍贯，班次，职务，固定电话，家庭住址）。

（4）主键（标识码）。主键是指表中的某个属性组，它可以唯一确定一个元组。例如表中的学号可以唯一确定一个学生，也就成为此关系的主键。

（5）域。属性的取值范围称为域。例如，学生的出生日期应该在公历日期规定的范围内，性别的域应该是（男，女），班级的域是学校该年级所有班次的集合。

（6）分量。元组中的一个属性值称为分量。

（7）关系模式。对关系的描述称为关系模式，一般表示为：

关系名（属性1，属性2，属性3，……）

例如，上面的关系模式可描述为：学生（姓名，学号，民族，出生日期，党/团时间，籍贯，班次，职务，固定电话，家庭住址），其中“学号”为主键。

4.2.4.2　关系模型的操作

关系模型中常用的关系操作主要包括查询和更新两类，其中查询操作包括选择、投影、连接、除、并、交、差等操作；更新操作包括插入、删除、修改等操作。

关系操作的基础是关系运算。关系运算可以归为两类：关系代数和关系演算。关系代数运算把关系当成集合，对它施加各种集合运算（如两个集合的并、差、交）和关系运算特有的投影、选择、求商、连接等运算。关系演算以数理逻辑中的谓词演算为基础，通过谓词形式来表示查询表达式。按照谓词变元的不同，关系演算又分为元组关系演算和域关系演算。

4.2.4.3　关系模型的完整性

在数据库中，数据的完整性指数据的精确性和可靠性。数据完整性包括以下两方面：

（1）与现实世界中的应用需求数据的相容性和正确性。

（2）数据库内数据之间的相容性和正确性。

例如，学生的学号必须是唯一的，性别只能是“男”或“女”，学生所选修的课程必须是已经开设的课程等。因此，数据库是否具有数据完整性特征关系到数据库系统能否真实地反映现实世界的情况。

数据完整性由完整性规则定义，而关系模型的完整性规则是对关系的某种约束条件。关系模型的完整性主要包括三类：实体完整性、参照完整性和用户定义完整性。数据库管理系统应该提供对这些数据完整性的支持。

1. 实体完整性

例如，在关系“学生（姓名，学号，民族，出生日期，党/团时间，籍贯，班次，职务，固定电话，家庭住址）”中，“学号”属性为主键，则该属性不能取空值。在关系“学生–课程（学号，课程号）”中，“学号”和“课程号”是主键，这两个属性值不能为空。

注意，上文中提到的空值（Null）不是0，不是空字符串，不是空格。空值表示没有值或“不知道”、“无意义”的值，是不确定的值。关系模型中的每一个元组都对应客观存在的一个实体，如一个学号唯一确定了一名学生，如果表中存在没有学号的学生数据，则该学生一定不属于正常管理范围的学生，甚至该学生是一个不存在的人。

2. 引用完整性

现实世界中的实体之间往往存在着某种联系。在关系模型中，实体与实体之间的联系都是用关系来表示的，这样就自然产生了关系与关系之间的引用。引用完整性就是描述实体之间的引用规则的。

例如，学生实体与成绩实体可以用下面的关系来表示，其中主键用下画线标识。

学生（姓名，学号，民族，出生日期，党/团时间，籍贯，班次，职务，固定电话，家庭住址）

成绩（学号，课程1成绩，课程2成绩，课程3成绩，……）

这两个关系存在着属性引用，成绩关系中的“学号”引用了学生关系中的主键“学号”。显然，成绩中的学号必须是确实存在的学生的学号，即在学校信息表中要有该学号。因此我们可以这样描述，成绩关系中的“学号”引用了“学生”关系中的“学号”的取值。

这样一个关系中某个属性的取值受到另一个关系中某个属性取值范围约束的特性就是引用完整性。

3. 用户定义的完整性

用户定义的完整性也称为域完整性或语义完整性。不同的关系数据库系统根据其应用环境的不同，往往还需要一些特殊的约束条件，用户定义的完整性就是针对某一具体关系数据库的约束条件。关系模型应提供定义和检验这类完整性的机制，以便系统用统一的方法处理。

例如，针对关系：学生（姓名，学号，民族，出生日期，党/团时间，籍贯，班次，职务，固定电话，家庭住址），依据客观世界的给定条件，此处的“班次”属性的取值范围为{1班，2班，3班，4班，5班}。

4.2.4.4 关系模型的范式

在设计数据库的时候，一个最根本的问题是怎样建立一个合理的数据库，使其无论是在数据存储方面，还是在数据操作方面都具有较好的性能，并且保证数据的完整性。为使数据库设计合理可靠、简单实用，长期以来已形成了关系数据库设计的规范化理论。它是根据现实世界存在的数据依赖而进行的关系模式的规范化处理，从而得到一个合理的数据库设计。

1. 规范化的提出

设有一个军校学生信息管理表，其中包括姓名、学号、党/团历史、党/团时间、固定电话、班次，班次地址等字段。按照一般建制规定，通常一个连可以下属有多个班，每个班可以有多个学员，部分数据如表4.4所示。

表4.4　学生信息管理表

姓名	学号	党/团历史	党/团时间	职务	班次编号	班次名称	固定电话	班次地址
潘越	3912006001	共青团员	1992.6.7	学员	1	4班	3513934	1-201
徐华	3912006002	共青团员	1992.6.7	班长	2	4班	3513934	1-201
白伟	3912006003	共青团员	1992.6.7	学员	2	5班	3513935	1-202
李思	3912006004	共青团员	1992.6.7	学员	2	5班	3513935	1-202

表4.4中存在许多重复数据，如每一行中的学生每分到一个相同班次，其班次地址、固定电话都会重复出现，引起重复存储；由于存在冗余数据，数据更新时可能会出现数据异常。如果更改某班电话，则需要修改多个元组。如果仅修改一部分数据，而其余部分数据未修改，就会造成数据的不一致性。

由于表4.4中各列之间存在着某种联系，如学生与班、班与班次地址、固定电话之间存在依赖关系，才使得数据出现大量冗余，引发各种操作异常。这种依赖关系称为数据依赖。为解决数据之间的依赖关系，常通过对表进行分解来消除不合理的部分，以减少冗余数据。可将表4.4分解为两个数据表，即学生基本信息表（学号，姓名，党/团历史，党/团时间，职务，班次）和班次信息表（班次编号，班次名称，固定电话，班次地址），分解后的部分数据如表4.5与表4.6所示。

表4.5　学生基本信息表

姓名	学号	党/团历史	党/团时间	班次名称
潘越	3912006001	共青团员	1992.6.7	4班
徐华	3912006002	共青团员	1992.6.7	4班
白伟	3912006003	共青团员	1992.6.7	5班
李思	3912006004	共青团员	1992.6.7	5班

表4.6　班次信息表

编号	班次名称	固定电话	班次地址
1	4班	3513934	1-201
2	5班	3513935	1-202

2. 规范化的含义

关系数据库中的关系必须满足一定的规范化要求，即可以用范式来衡量不同的规范化程度。范式是符合某一种级别的关系模式的集合，是衡量关系模式规范化程度的标准，满足最低要求的范式称为第一范式，在第一范式基础上进一步满足一些要求的为第二范式，其余以此类推。

关系模式的规范化主要用来解决关系中冗余数据及由此产生的操作异常，也是消除关系模式中产生数据冗余的数据依赖。

3. 第一范式

如果关系模式R的每个关系r的属性值都是不可分的原子项，则称R满足第一范式的模式。我们将满足第一范式的关系称为规范化的关系，不满足第一范式的称为非规范化的关系。第一范式是关系模式应具备的最基本的条件。

(1) 示例一

判断表4.7所示的关系模式是否符合第一范式的基本要求。

表4.7 示例一的数据表

姓名	学号	党/团历史	党/团时间	职务	固定电话	班次	班次地址
潘越	3912006001	共青团员	1992.6.7	学员	3513934	4班	1-201
徐华	3912006002	共青团员	1992.6.7	班长	3513935	5班	1-202
		中共党员	2010.6.7				
白伟	3912006004	共青团员	1992.6.7	学员	3513936	6班	1-203

依据第一范式的要求，表“党/团历史”属性的值可再分为“共青团员”与“中共党员”，因此不符合第一范式，可将表4.7改进为表4.8所示。在表4.8中，由“学号”和“党/团历史”共同组成主属性，区分每一个学员。

表4.8 对示例一修正后的数据表

姓名	学号	党/团历史	党/团时间	职务	固定电话	班次	班次地址
潘越	3912006001	共青团员	1992.6.7	学员	3513934	4班	1-201
徐华	3912006002	共青团员	1992.6.7	班长	3513935	5班	1-202
徐华	3912006002	中共党员	2010.6.7	学员	3513935	5班	1-202
白伟	3912006004	共青团员	1992.6.7	学员	3513936	6班	1-203

（2）示例二

判断表4.9所示的关系模式是否符合第一范式的基本要求。

表4.9 示例二的数据表

学号	姓名	成绩	
		计算机网络	数据结构
3912006001	潘越	92	88
3912006002	徐华	80	93
3912006003	倪涛	82	76
3912006004	白伟	72	80

依据第一范式的要求，表4.9所代表的关系模型不符合第一范式要求，由于“成绩”属性值又分为了“计算机网络”成绩与“数据结构”成绩，而不是不可分的原子项，因此需要将表进行规范化处理，即可变为表4.10所示。

表4.10 对示例二修正后的数据表

学号	姓名	计算机网络	数据结构
3912006001	潘越	92	88
3912006002	徐华	80	93
3912006003	倪涛	82	76
3912006004	白伟	72	80

4. 第二范式

对于满足第一范式的关系模式R，如果所有非主属性都完全依赖于候选关键字，则称R属于第二范式，简而言之，第二范式就是非主属性非部分依赖于主关键字。

第二范式是在第一范式的基础上建立起来的，即满足第二范式必须先满足第一范式。第二范式要求实体的属性完全依赖于主关键字，即不能存在仅依赖主关键字一部分的属性。如果存在，那么这个属性和主关键字的这一部分应该分离出来形成一个新的实体，新实体与原实体之间是一对多的关系。

例如，在表4.8中，为了区分每条记录，主键定义为“学号”和“党/团历史”两项，用于共同区分每条记录。但是，不难发现该表中存在数据冗余，即姓名为徐华的记录中的“固定电话”、“班次”和“班次地址”三项具有相同的数据值，造成了冗余，它不满足第二范式。“固定电话”、“班次”和“班次地址”依赖于部分主关键字，即只依赖于主属性中的“学号”字段，而与“党/团历史”无关，这违反了第二范式中的“非主属性非部分依赖于主关键字”的原则。

为了改进以上数据结构，使其符合第二范式，可将表4.8改成以下表4.11和表4.12所示。

表4.11　修正的学员信息表

姓名	学号	职务	固定电话	班次	班次地址
潘越	3912006001	学员	3513934	4班	1-201
徐华	3912006002	班长	3513935	5班	1-202
徐华	3912006002	学员	3513935	5班	1-202
白伟	3912006004	学员	3513936	6班	1-203

表4.12　修正的学员党/团历史情况表

学号	党/团历史	党/团时间
3912006001	共青团员	1992.6.7
3912006002	共青团员	1992.6.7
3912006002	中共党员	2010.6.7
3912006004	共青团员	1992.6.7

5. 第三范式

关系模式R(U, F)中若不存在这样的键X、属性组Y及非主属性Z（$Z \notin Y$），使得$X \to Y$，$Y \not\to X$，$Y \to Z$成立，则称R(U, F)属于第三范式。其中，$X \to Y$指的是Y函数依赖于X，$Y \not\to X$指的是X函数不依赖于Y，$Y \to Z$指的是Z函数依赖于Y。

表4.13　原始学员信息表

姓名	学号	固定电话	班次	班次地址
潘越	3912006001	3513934	4班	1-201
徐华	3912006002	3513935	5班	1-202
白伟	3912006004	3513936	6班	1-203
于成庆	3912006005	3513936	6班	1-203

例如，在表4.13中，于成庆与白伟均在6班，因此“固定电话”与“班次地址”项的数值出现重复，会造成一定的冗余，该关系模式不符合第三范式的要求。

在该关系模式中，“学号”属性为键，对应第三范式定义中的键X，非主属性“姓名”和“班次”对应于定义中的属性组Y，“固定电话”和“班次地址”字段对应于定义中的非主属性Z。经过分析发现，“姓名”和“班次”属性依赖于主键“学号”字段，即满足“$X \to Y$”，但是“学号”属性不依赖于“班次”和“姓名”属性，即亦满足“$Y \not\to X$”，且“固定电话”和“班次地址”依赖于“班次”属性，即满足“$Y \to Z$”，因此在该关系模式中存在这样的键X、属性组Y及非主属性$Z(Z \notin Y)$，使得$X \to Y$，$Y \not\to X$，$Y \to Z$成立，从而违反了第三范式的原则。因此需要加以改进，即将上述的关系模式改为如下两个关系模式，如表4.14和表4.15所示。

表4.14 修正的学员信息表

学号	姓名	班次
3912006001	潘越	4班
3912006002	徐华	5班
3912006004	白伟	6班
3912006005	于成庆	6班

表4.15 修正的班次信息表

编号	班次	固定电话	班次地址
1	4班	3513934	1-201
2	5班	3513935	1-202
3	6班	3513936	1-203

改进后的关系模式不再出现非主属性传递依赖于主属性的问题，因此也符合了第三范式的要求。

4.2.5 物理模型

物理模型是面向计算机物理表示的模型，描述了数据在存储介质上的组织结构，它不但与具体的数据库管理系统软件有关，而且还与操作系统和硬件有关。该模型用于存储结构和访问机制的更高层描述，描述数据是如何在计算机中存储的，如何表达记录结构、记录顺序和访问路径等信息。数据库的物理设计阶段必须在此基础上进行详细的后台设计，设计内容包括数据库的存储过程、操作、触发、视图和索引等。

4.3 数据仓库技术

4.3.1 数据仓库的起源

随着信息技术的发展和应用，人们生产和收集数据的能力大幅提升，无数个数据库被用于商业管理、政府办公、科学研究和工程开发等，这一势头仍将持续发展下去。但是，一个新的挑战被提出了：在这个“信息爆炸”的时代，信息过量几乎成为人人需要面对的问题，如何更好地利用大量数据，从中及时发现有用的知识，提高信息利用率，进而来应对“数据丰富而知识贫乏”的挑战呢?

数据库技术在经过了20世纪80年代的辉煌之后，已经在各行各业成为一种文化或时尚，数据库界目前除了关注分布式数据库、面向对象数据库、多媒体数据库、查询优化和并行计算等技术外，已经开始反思：数据库实质的应用不应仅仅局限于查询，更应侧重于发现，进而更好地为管理者决策服务。

因此，能否从纷繁复杂、大量沉淀的数据环境中得到有用的决策信息，及时做出正确的分析与决策，已成为企业生存与发展的至关重要的环节。自从20世纪70年代提出决策支持的概念以来，人们在决策支持系统的理论及应用上做了大量的研究工作，并在企业决策中发挥了积极的作用。但是，现有的大部分决策支持系统是基于传统数据库基础之上的，随着企业的数据量的不断增加，需要对原有的信息进行提炼和加工，需要为领导提供集成化和历史化的数据，需要为企业全局的战略决策和长期趋势分析提供更有效的支持。在这种情况下，传统的数据库管

理系统因自身的局限性已无法满足决策支持系统对数据的要求。因此，一种适用于决策支持系统的数据组织与管理技术——数据仓库技术应运而生，并逐渐成为支持分析与决策的重要技术。

4.3.2　数据仓库的基本特征

数据仓库与事务性数据库系统相比，有着本质的区别。事务性数据库是一种通用的建立在严格的数学模型之上的系统，用来管理企业数据，进行事务处理，完成相关业务。而数据仓库没有严格的数据理论，更偏向于工程，它不是花钱就能购买到的，而是数据日积月累的建立过程。它的应用对象是不同层次的管理者。数据仓库有多种数据源，库中数据一般无须修改删除，主要用于进行大规模查询和分析操作，因此具有大量的历史数据和汇总数据。一般数据仓库具有以下四个特征。

1. 数据仓库是面向主题的

事务性数据库面向操作，侧重于细节，而数据仓库面向主题，更侧重于高层次的聚合。例如，购买最流行儿童书籍的家长不会去关注青少年小说产品的库存信息，但计划重新安排书店的图书摆放形式的管理员可能会对“计算机书籍的销售主题”感兴趣，可根据销售情况重新安排图书摆放形式。

2. 数据仓库中的数据是集成的

事务性数据库通常是为特定的应用程序设计的，然而数据仓库集成了不同来源的数据。例如，订单处理应用程序及其数据库可能会包含每个订单的详细折扣信息，却没有任何生产成本警戒信息。相反，生产应用程序及其数据库可能包含了详细的成本信息，却没有任何销售折扣信息。通过把这两个数据源集成到数据仓库中，可以计算出产品销售的实际利润。

3. 数据仓库中的数据是相对稳定的

事务性数据库关心的是现在，数据仓库关心的是随着时间变化的活动。例如，对于一个简单的银行账户，每个交易即每笔存款或取现，都会对账户余额产生瞬间改变，但事务数据库很少维护历史余额。而在数据仓库中可以存储好几年的交易数据，以便将当前数据与上月或上一年的数据进行对比。制定决策时，能够看到很长的时间范围，这对区分趋势和随机波动很重要。

4. 数据仓库中的数据是随时间不断变化的

数据仓库中的数据不是一成不变的，而是随时间变化的。需要不断获取联机事务处理系统的不同时刻的数据，将其集成后追加到数据仓库中，因此数据仓库中数据的键都包含时间项，以表明数据的历史时期，并可在时间维度上对数据进行分析。同时，数据仓库中的数据也有时间期限，在新数据不断进入的同时，过时的数据也要从数据仓库中排除出去。

4.3.3　数据仓库的相关概念

为了更好地理解数据仓库的概念和进一步讨论数据仓库的具体技术，下面详细介绍几个数据仓库的相关概念。

1. 粒度

粒度是对数据仓库中的数据综合程度高低的度量。粒度既影响数据仓库中数据量的规模，也影响数据仓库所能回答询问的种类。粒度越小，综合程度越低，回答查询的种类越多；而粒度越大，综合程度越高，查询的效率也越高。

2. 维度

维度是一个物理特性（如时间、地点、产品等），是表达数据仓库中信息的基本途径，可作为标识数据的索引。通常报表只包含有行和列两维，但在数据仓库中存储的数据大多是用多维（三维或三维以上）视图表示的。

3. 聚合

在数据仓库技术中，每一维度可包括多个层次，这些层次可以向用户提供某一层次的数据。例如，在地理位置维中，由所有的街区组成了地区，由所有的地区组成了城市等。聚合就是指把最细粒度的事实数据按照维度的不同层次进行汇总，从而构成维度内不同层次的数据集，使用户不仅能够在一个维度内观察数据，而且能够在维度内的不同层次上观察数据。

4. 分解与合成

分解与合成是指在一个维度内进一步细分数据，或将数据按照另一标准进行组合的过程。例如，当以地理位置维度观察数据时，用户可以首先以国家（如中国）为单位观察数据，然后可以选择观察某一个地区（如华北地区）的数据，接下来可以观察某一个省或城市（如石家庄）的数据，这就是数据分解的过程。而合成则是分解的逆过程，例如用户开始以省、市为观察对象，接着再以地区、国家等为观察对象，就是一个数据合成的过程。

4.3.4 数据仓库的体系结构

数据仓库从多个数据源或者异质数据源获取原始数据，经过整理加工后，存储起来，通过向用户提供各种访问工具，提供统一、协调和集成的信息环境，以支持企业的决策工作。一般而言，数据仓库系统的体系结构包含五个层次，分别是：数据源，数据提取、加载、清理和刷新，数据存储与管理，OLAP服务器和前端工具等（见图4.10）。在该体系结构中主要涵盖了以下几个主要模块。

1. 数据源。数据源是数据仓库的基础，通常包括外部数据、业务数据及各类文档等。

2. ETL（Extract Transform Load，数据的提取、转换和加载）。该模块负责从外部数据源获取数据，进行转换并准备装入数据仓库。这个模块作为数据挖掘或商业智能软件的核心和灵魂，能够按照统一的规则集成并提高数据的价值，完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。如果说数据仓库的模型设计是一座大厦的设计蓝图，数据是砖瓦，那么ETL就是建设大厦的过程。根据国内外众多实践中的普遍共识，认为ETL的规则设计和实施工作量最大，约占整个项目的60%~80%。

3. 数据存储与管理。该模块建立整个数据仓库的核心，负责数据仓库的内部维护和管理，包括数据安全、归档、备份、维护和恢复等功能。在现有系统的基础之上，使用ETL并有效集成，按照主题进行重新组织，最终确定数据仓库的存储结构，同时组织存储数据仓库元数据。根据实际情况及企业部署数据仓库的经验，数据仓库可以分为企业级数据仓库和部门级数据仓库（数据集市）。

4. OLAP服务器。OLAP就是联机分析处理，它是一类软件技术，可以使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的、并真实反映企业维度特性的信息进行快速、一致、交互地存取，从而获得对数据的更深入了解。简单地说，就是对数据做多维分析。OLAP服务器就是用来提供这种支持的。

5. 前端工具与应用。前端工具主要包括各种报表工具、数据挖掘工具、查询工具，以及各种基于数据仓库或数据集市的开发应用。这个阶段面向最终用户，因此界面必须简单易用且高效，功能较为强大。考虑数据的安全性，必须具有良好的权限控制。

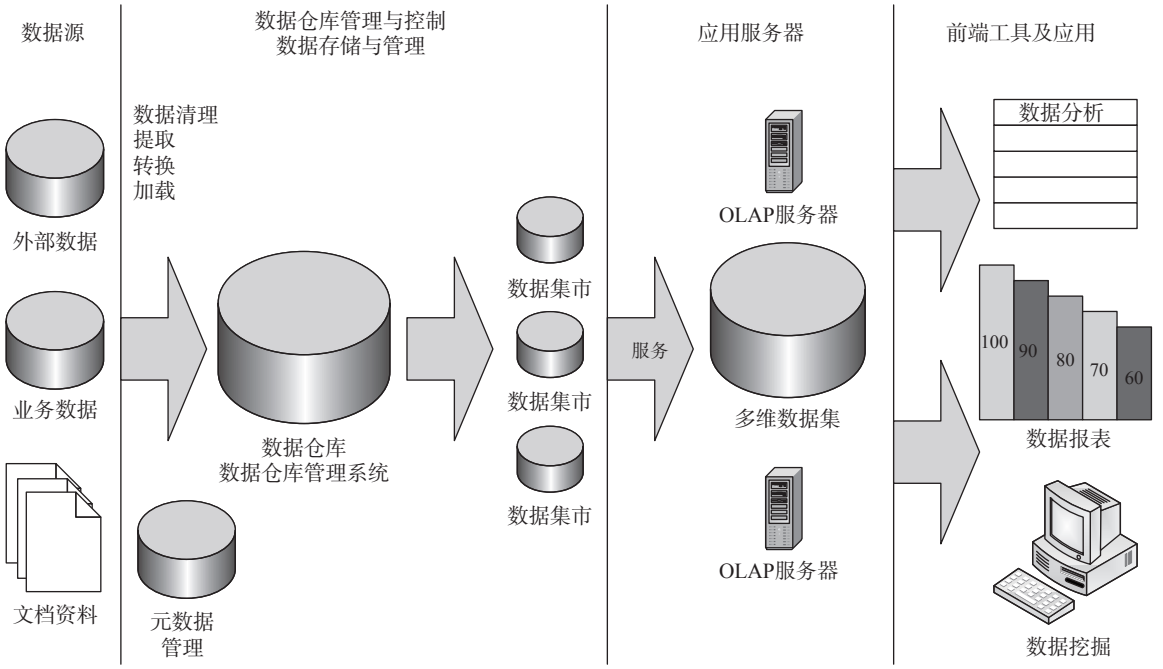


图4.10　数据仓库的体系结构图

4.4　存储技术

计算机技术的出现与发展促使信息存储技术经历了第二次质的变革。通过计算机信息技术，人们可以很好地解决先前的信息存储技术存在的一系列问题。从此，信息存储领域开始真正进入成熟期，其发展速度日新月异，新的信息存储理念和技术不断推出。信息存储实际上是信息系统的一部分，它和计算机系统与技术的联系非常紧密。本节以存储系统结构为主，介绍几种常见的存储技术。

4.4.1　RAID基础知识

军事信息系统涉及海量信息的存储、处理和访问，数据的存储、备份与恢复尤为重要。

在较为恶劣的环境下存储数据时，单盘系统的安全性存在一定的隐患。为此，在数据中心或重要的计算机上，常使用RAID（独立磁盘冗余阵列）作为存储设备。RAID采用多盘方式存储数据，同时通过并行技术和数据保护技术，提高系统的容量、速度和可靠性。

根据RAID系统对磁盘数据分布以及校验方式的不同，RAID系统可以分为7个级别（RAID0~RAID6）。其中RAID0、RAID1、RAID3和RAID5是较为常用的级别。

RAID0

RAID0是无冗余、无校验的磁盘阵列，又称为条带化的阵列，全称是“没有容错设计的条带磁盘阵列”。RAID0的工作原理如图4.11所示。数据在存储时，由RAID控制器将其分割成大小

相同的数据条，并写入阵列中的各个硬盘。我们可以将分块存入各硬盘的数据抽象成一个数据条带横跨阵列中的所有硬盘，至于条带的宽度要根据具体的系统而定。在读取数据时，也同时从各硬盘中读取条带块。由于读写基本上都是基于并行的数据传输，所以这种阵列模式的读写速度在所有RAID类型中最快。实际上，RAID0的数据是顺序传输的，但是系统将多个硬盘读写操作相互重叠执行，使得RAID0的数据读写速度要远远大于分别读写这些硬盘的速度。

但是，RAID0的主要缺点就是没有容错能力。它将数据分段写入阵列中的多个硬盘中，若其中任何一个硬盘出现问题，都会导致所存入的数据部分损坏，而部分数据的丢失将导致整个数据的失效。另外，RAID0只是简单地将多个硬盘进行组合而形成的阵列，整个RAID0系统的安全性反而还不如单个硬盘高，所以RAID0只适合于对数据安全性要求不高但对速度要求很高的场合。

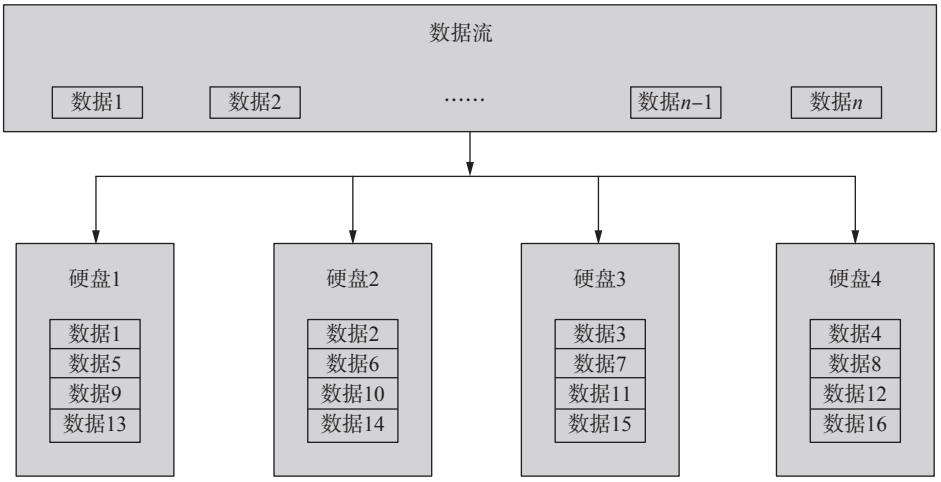


图4.11 RAID0 工作原理

RAID1

虽然 RAID0可以提供更大的容量和更快的读写速度，但是没有冗余功能。如果出现故障，系统将停止运行。RAID1和RAID0截然不同，其技术重点在于如何能够在不影响性能的情况下，最大限度地保证系统的可靠性和可修复性。

RAID1又称为镜像阵列，它将同样的数据写入两块硬盘，两块硬盘互为镜像盘。当一块硬盘中的数据受损或磁盘故障时，另一块硬盘可继续工作，并可在条件具备时重建 RAID1阵列（见图4.12）。

然而，RAID1不能提升磁盘性能，两块硬盘组成RAID1后的容量等于单块硬盘的容量，对任何一个磁盘的数据写入都会被复制到镜像盘中，系统可以从一组镜像盘中的任何一个磁盘读取数据，因此磁盘镜像明显提高了系统的成本。RAID1适合对数据可靠性要求严格的场合，比如金融、保险、交通等重要部门以及网络服务器等，用来保存关键性的重要数据。

在RAID1中，如果一块硬盘失效，系统就会忽略该硬盘，转而使用镜像盘读写数据。通常，把出现硬盘故障的RAID系统称为在降级模式下运行。虽然这时保存的数据仍然可以继续使用，但是RAID系统将不再可靠。如果剩余的镜像盘也出现问题，那么整个系统就会崩溃。因此，应当及时更换损坏的硬盘，以避免出现新的问题。更换新盘之后，原有的数据必须复制到新盘中，这一操作称为同步镜像。同步镜像一般都需要较长时间。在同步镜像的过程中，外界

对数据的访问不会受到影响，但由于复制数据需要占用一部分网络带宽，所以可能会使整个系统的性能有所下降。

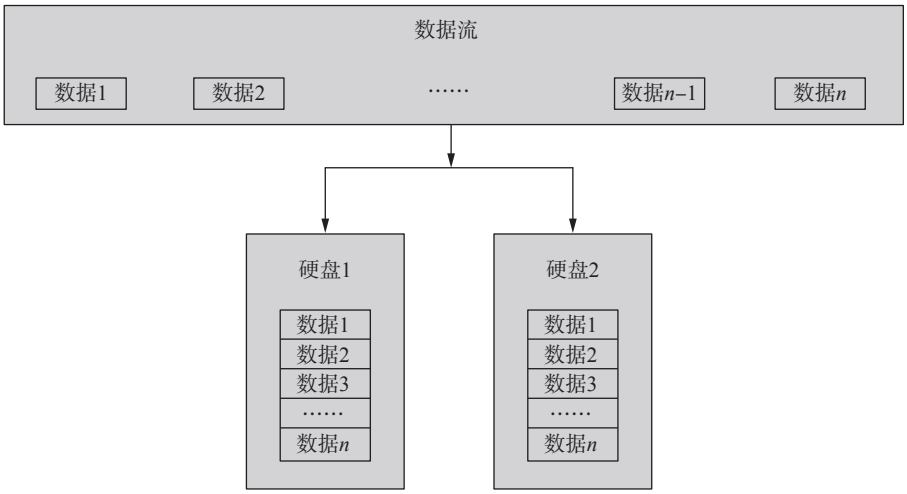


图4.12　RAID1工作原理

RAID2

RAID2又称为纠错汉明码磁盘阵列。它是目前RAID阵列中最复杂的级别，其主要原因在于它所采用的错误检测与修复技术为汉明码。汉明码是在原有数据中插入若干校验码来进行错误检查和纠正的编码技术，根据现有的数据编码各位的值，生成校验码，然后再将校验码与原有的数据码合并，转换成新的数据编码。

在RAID2中，硬盘数量是根据数据存储位宽决定的，其个数等于$2^P = P + D + 1$（P为汉明码的个数，D是原始数据的位数）。例如，对于4位数据宽度，就需要4个数据盘和4个汉明码盘，$2^3 = 3 + 4 + 1$。图4.13就是以校验值个数为3显示的RAID2逻辑结构图。

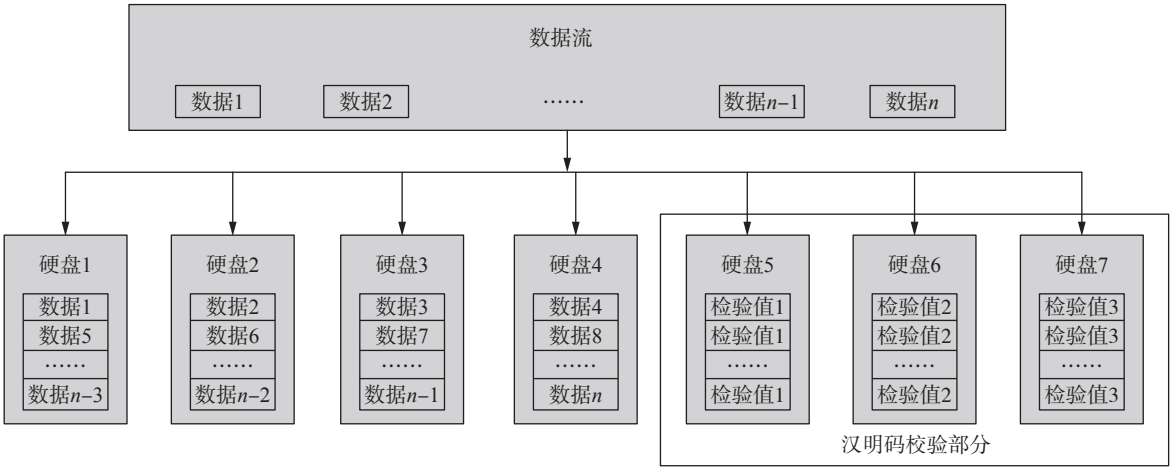


图4.13　RAID2工作原理

根据汉明码在原数据码中插入的位置，对应到阵列中硬盘的相应位置，以第1个、第2个、第4个、…、第2n个硬盘的顺序，按一定位数取出作为专门的校验盘，用于校验和纠错。例如，RAID2是5位数据宽度时，其中数据盘为5块，校验盘为4块，纠错盘占总盘数的40%左右；而当

RAID2是64位宽度时，则数据盘为64块，校验盘为7块，纠错盘占总盘数的近10%。所以，使用硬盘数越多，则纠错盘在其中所占的空间比越小。

RAID2是早期为了能够进行即时数据校验而发明的校验阵列。在读写过程中，该阵列都要求计算数据相关的汉明码并进行校验。汉明码只能纠正一位错误，也只允许一个硬盘出错。RAID2对大数据量的读写具有很高的性能，但对少量数据的读写性能不好。在实际应用中，由于其存储空间利用率、读写性能和花费成本等方面的问题，使得RAID2很少使用。

RAID3

针对RAID2的缺点，研究人员设计出一种新的更为先进的数据纠错阵列：RAID3，也就是并行传输及校验阵列。它使用相对简单的异或运算校验代替了较为复杂的汉明码校验，大大降低了阵列的实现成本。异或运算是位运算，其运算规则是：两个数据位异或，当两个数据位相同时结果为0，不同时结果为1。它有一个特殊的性质，如果a异或b等于c，那么a异或c也等于b，b异或c也等于a。当然，这种性质可以推广到任意多个运算数的情况，这对于数据恢复非常有用。

假设某一RAID3阵列由5块硬盘构成，分别编号为1~5，其中5号盘为校验盘，1~4号盘为数据盘，则校验盘中的数据可通过各块数据盘中的对应位数据进行异或运算获得。具体计算公式如下：

$$D_5(i) = D_4(i) \oplus D_3(i) \oplus D_2(i) \oplus D_1(i)$$

如果在某个时刻得知1号盘中的数据发生了故障，则可依据RAID3的校验原理对1号盘数据进行恢复，计算公式如下：

$$D_1(i) = D_5(i) \oplus D_4(i) \oplus D_3(i) \oplus D_2(i)$$

如图4.14所示，RAID3也是采用数据分块并行传送的方式读写数据的，它先将数据分块，再计算它们的异或校验码，然后把分块数据和异或校验码一起写入阵列中。采用这种方法可以提高数据的存取速度和可靠性，而且当任一硬盘出现故障时，都可利用剩余硬盘的数据重构失效盘上的数据。RAID3的硬盘利用率比RAID1要高，但由于异或校验的信息固定存储在一个硬盘上，可能会造成该校验盘负荷过重，从而影响阵列的整体性能。RAID3在执行大量数据写操作时，校验盘会成为瓶颈，因此RAID3更适用于写操作较少、读取操作较多的应用环境。

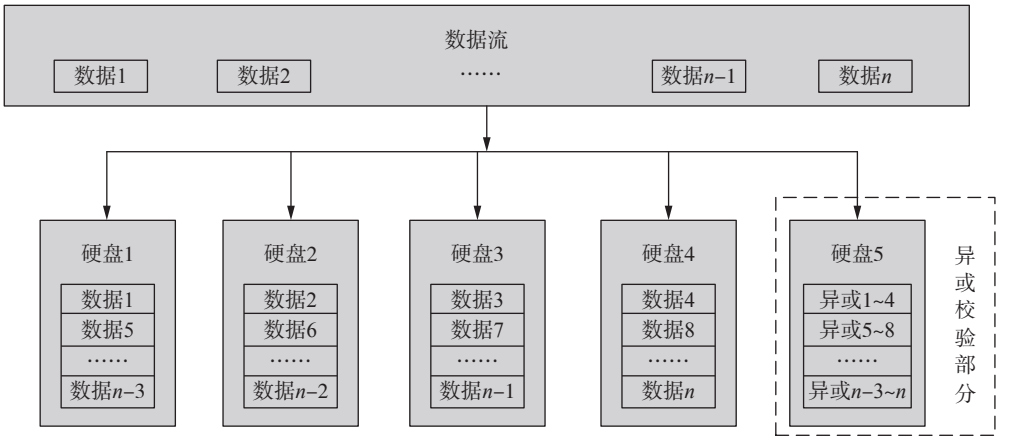


图4.14 RAID3工作原理

RAID4

RAID4是独立的数据盘与共享的校验盘阵列技术，其关键之处在于充分利用数据块的优点，它在分配存储空间时是以数据块为单位进行的。RAID4将数据块划分为足够包含一个完整的文件，其主要目的在于保证存储数据的局部完整性，使其不会在对完整数据进行条带化分散存储后，出现因多个硬盘出错而对数据产生影响的局面。这一想法虽好，但却无法真正实现。

实际的RAID4是一种块交叉异或校验冗余的阵列，它与RAID3一样，仍采用一个专用的校验盘来存储相关的校验数据。RAID4主要采用了要被替换的旧数据、新写入的数据和旧的校验数据来计算新的校验数据的方法。也就是说，RAID4只需要读写有用的数据盘和校验盘，而不会像RAID3那样每一次写操作都要访问所有硬盘。在写入时，RAID4就是按这个方法把各硬盘上对应数据的校验值统一写入校验盘。当某个硬盘故障时，可以通过剩余硬盘上的对应数据恢复故障盘数据。

由于RAID4在写入时要顺序写硬盘，同时还要写校验数据，所以写效率比较差；读取时虽然也是逐个硬盘顺序读取，但总体速度比较快，所以读性能要强一些。RAID4的结构图基本上与RAID3相似。

RAID5

RAID5是目前市场上最常见的RAID产品，它是一种无独立校验盘的奇偶校验硬盘阵列，具有较高的存储性能、较强的数据安全性以及较低的实现成本等特点。RAID5同样也是采用异或校验来检查阵列中的存取错误的，但是它没有独立的校验盘，而是把数据与其对应的奇偶检验信息分别存放到不同的硬盘上。当某块硬盘出现故障时，可以利用剩下的硬盘数据来恢复故障盘上的数据。

RAID5的交叉写盘技术主要用来减少阵列写操作时的瓶颈问题。数据的写入过程如图4.15所示，从中也可以看到数据与校验值的分布情况。

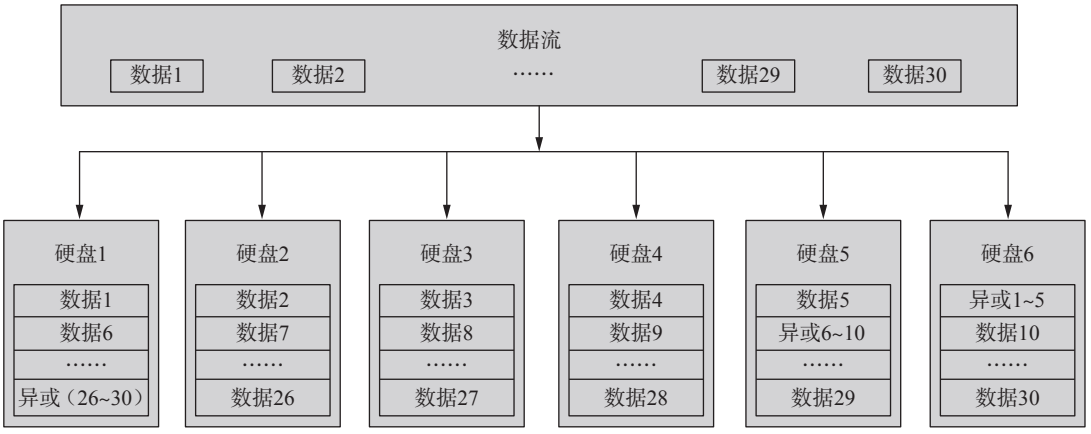


图4.15　RAID5工作原理

RAID5为数据提供的数据安全保障级别比镜像低，但存储空间的利用率却要比镜像高。RAID5具有和RAID0相近的数据读取速度，只是多了一个奇偶校验信息。但它的写入速度较低，存在“写损失”的现象，尤其是写许多小文件和随机数据时更明显。每一次写操作将产生4个实际的读/写操作，其中2次读旧的数据及奇偶信息，2次写新的数据和奇偶信息。

在成本、可靠性和性能都同等重要时，可以采用RAID5。事实上，RAID5被广泛应用于各种类型的服务器，如文件服务器、应用程序服务器、数据库服务器、Web服务器和E-mail服务器等。

RAID6

早期磁盘阵列设计的主要目的是防止某单个硬盘失效而影响系统正常运行，也就是说，一般情况下，若两个以上的硬盘出错，则阵列的数据将无法恢复。RAID6是一种双重奇偶校验存取阵列，在设计时引入了两级冗余技术，主要目的是保证阵列在两个硬盘同时失效时仍能够正常工作，并恢复相应的数据。

如图4.16所示，用P0~P5分别代表硬盘上条带区的奇偶校验值，用QA~QF分别代表硬盘A~F中数据块的校验值，它的实现方式就是进行两次独立的校验计算，并把校验值分别存放在两个校验盘上（也可以交叉存放校验值）。这时，若两个硬盘失效，则能够通过求解带两个变量的方程组来恢复失效盘的数据。RAID6控制器的设计非常复杂，实现成本极高，目前主要用于数据绝对安全的环境中，不太适于普及使用。

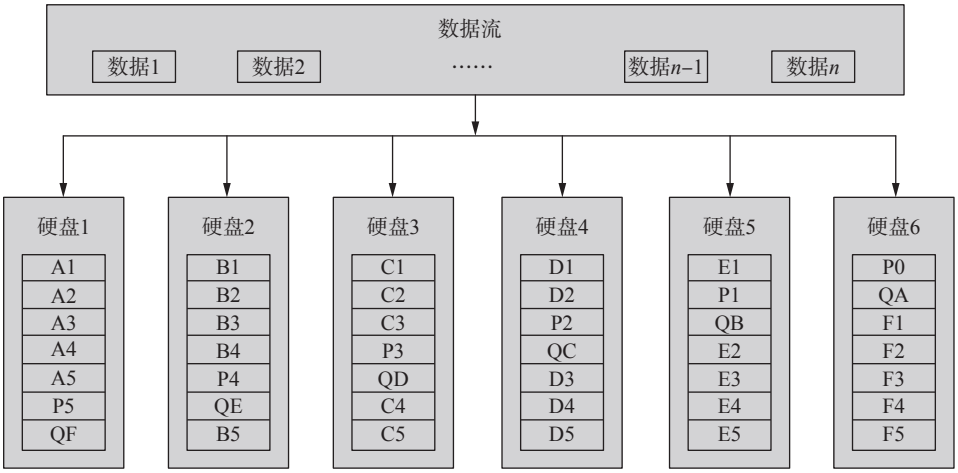


图4.16 RAID6工作原理

RAID0+1

RAID0+1结合了RAID0和RAID1的优点，采用两组RAID0的磁盘阵列互为镜像，也就是它们又构成一个RAID1阵列。在每次写入数据时，磁盘阵列控制器会将数据同时写入两组“大容量阵列磁盘组（RAID0）”中。数据除分布在多个磁盘上以外，在每个磁盘还都有其物理镜像盘，提供全冗余能力，允许一个磁盘故障而不影响数据可用性，并具有快速读/写能力。

至少4个磁盘才能做成RAID0+1。如果是磁盘A、B、C、D部署RAID0+1方案，可以使用磁盘A、C部署为RAID0，使用磁盘B、D部署为RAID0，然后将两个RAID0 部署为RAID1（见图4.17）。RAID0+1以高成本为代价换取存储性能和数据安全兼顾的方案，在提供与RAID1一样的数据安全保障的同时，也提供了与RAID0近似的存储性能。其最大容量为：磁盘数×磁盘容量/2。

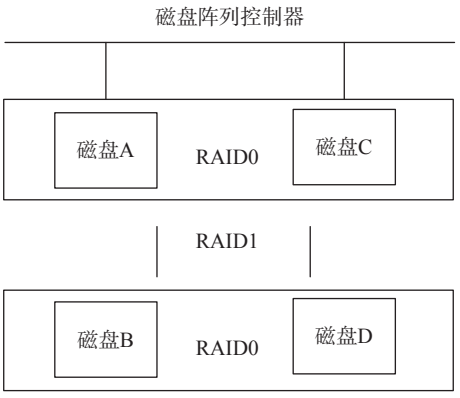


图4.17 RAID0+1工作原理

从理论上讲，RAID0+1能够经受住RAID0阵列中任何一块磁盘的故障。很多RAID0控制器会在RAID0阵列中的某一块磁盘出现故障之后让RAID0离线，这时候只有剩下的RAID0阵列在工作，因此系统将不存在冗余功能。简而言之，如果每个RAID0阵列中都有一块磁盘出现故障，那么整个磁盘阵列将不再工作。

RAID0+1适用于既有大量数据需要存取，同时又对数据安全性要求严格的领域，例如银行、金融、档案管理等。

RAID1+0

RAID1+0是由多组RAID1构成的RAID0，它的磁盘空间利用率与RAID0+1的相同。

RAID1+0部署结构（以4个磁盘为例）如图4.18所示。4块物理磁盘分别定义为A盘、B盘、C盘、D盘。先建立两组RAID1，第一组RAID1由A和B组成，命名为AB，第二组RAID1由C和D组成，命名为CD，然后再由AB和CD组成RAID0。

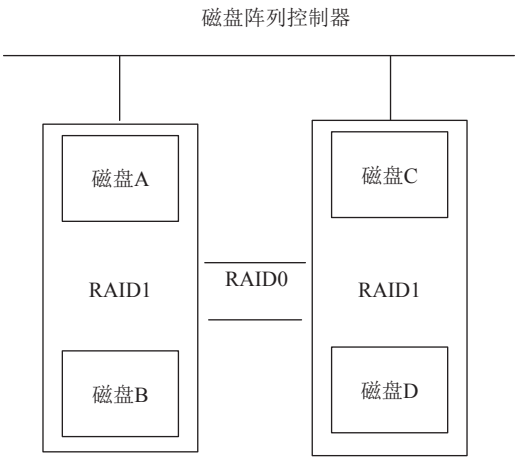


图4.18　RAID1+0工作原理

RAID0+1与RAID1+0虽然都利用了RAID0和RAID1的优点，但两种架构的安全性和性能有较大区别。

在数据安全性方面，在RAID0+1中，若有一块磁盘（假设为A盘）出现故障，则A盘所在的AB组RAID0也不再工作，只剩下CD一组RAID0提供服务，此时系统失去镜像功能，安全性等同于RAID0；而在RAID1+0中，若同样有一块磁盘（假设为A盘）出现故障，则除A盘以外的其他磁盘仍正常提供服务，虽然可靠性有所降低，但比第一种情况下的RAID0强得多。当组成RAID的磁盘数量增加时，这种可靠性的差距会更大。

在读写性能方面，两者的读写性能几乎没有差异。

RAID0+1与RAID1+0相比，唯一的好处是组成两个RAID0的磁盘个数和容量可以不一致，而RAID1+0则要求所有的磁盘容量完全一致。

1. RAID各级别性能比较

在实际工作中，选择相应的RAID级别主要根据3个因素，即可用性、存储容量和实现成本。如果不要求可用性，选择RAID0就可以获得最佳存储性能。如果可用性和存储容量很重要而实现成本无须考虑，则可以根据硬盘数量选择RAID1或RAID6。如果可用性、存储容量和实现成本都需要考虑，则可根据具体的应用环境和情况选择RAID3或RAID5。不同级别的RAID在性能方面的差异如表4.16所示。

表4.16　RAID各级别性能比较

RAID级别	磁盘数目	容量	存储效率	容错性	可用性	随机读性能	随机写性能	顺序读性能	顺序写性能	成本
RAID0	≥2	$S\times N$	100%	无	★	★★ ★★	★★ ★★	★★ ★★	★★ ★★	￥
RAID1	2	$S\times N/2$	50%	★★ ★★	★★ ★★	★★ ★	★★	★★	★★★	￥￥￥

（续表）

RAID级别	磁盘数目	容量	存储效率	容错性	可用性	随机读性能	随机写性能	顺序读性能	顺序写性能	成本
RAID2	不定	不定	70%~80%	★★	★★	★★	★	★★	★★★	￥￥￥
					★★			★★		￥￥
RAID3	≥3	$S\times(N-1)$	$(N-1)/N$	★★	★★	★★	★	★★	★★★	￥￥
				★	★★	★		★★		
RAID4	≥3	$S\times(N-1)$	$(N-1)/N$	★★	★★	★★	★★	★★★	★★	￥￥
				★	★★	★★				
RAID5	≥3	$S\times(N-1)$	$(N-1)/N$	★★	★★	★★	★★	★★	★★★	￥￥
				★	★★	★★★		★★		
RAID6	≥42	$S\times(N-2)$	$(N-2)/N$	★★	★★	★★	★	★★	★★	￥￥
				★★★	★★★	★★★		★★		￥￥

4.4.2 网络存储技术

网络上用于共享的存储设备通常都使用RAID，存储设备的连接和组织方式称为网络存储结构。网络存储结构通常分为三种：直连式存储（Direct Attached Storage，DAS）、网络附加存储（Network Attached Storage，NAS）和存储区域网络（Storage Area Network，SAN）。

4.4.2.1 直连式存储

在直连式（DAS）结构中，存储设备通过电缆（通常是SCSI接口电缆）直接连接服务器，服务器的I/O请求直接发送到存储设备。存储设备依赖于服务器，其本身是硬件的堆叠，不带有任何存储操作系统。

DAS的典型结构如图4.19所示。其中的存储设备一般都是RAID。

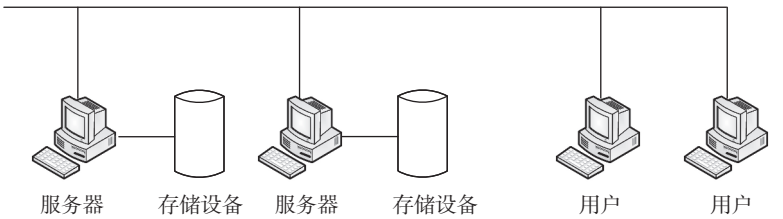


图4.19 DAS典型结构

DAS是最早在网络中采用的存储结构，具有存取速度快，建立方便等优点，但是它也有一些明显的问题。

1. 单点故障问题

由于一般都采取了冗余措施，所以存储设备的可靠性比较高。网络中的存储设备是共享的，在DAS结构下，其他所有设备只能通过连接存储设备的服务器使用存储设备。通常，服务器的可靠性低于存储设备，当连接存储设备的服务器发生故障时，其他所有设备都将无法使用存储设备，导致整个网络无法正常工作。克服单点故障的措施是使多个服务器共享一个存储设备，形成如图4.20所示的直连式共享存储系统。在这种结构下，当某台服务器出现故障时，用户可以通过另一台服务器访问存储设备，仍然可以正常工作。

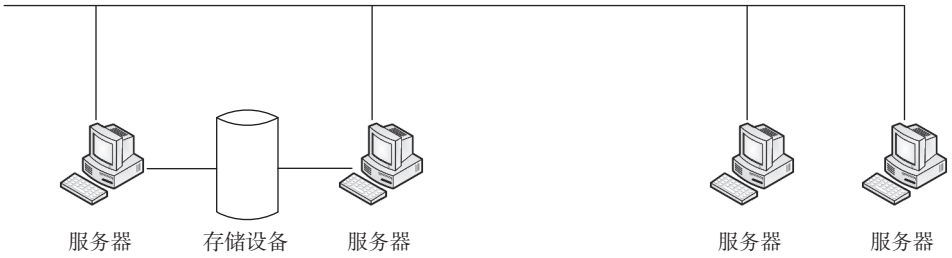


图4.20　直连式共享存储系统的存储结构

2. 扩展困难

每台RAID设备能容纳的磁盘数是固定的，一般在购买时不会留太多余量。当存储设备容量不够时，不能增加磁盘，就需要增加存储设备。而服务器可挂接存储设备的接口也是有限的，只能挂接有限数量的存储设备。当所需要的存储容量超过挂接能力时，系统就很难再继续扩展，或者添加设备需要较高的费用。同时，添加设备后，容易出现所有服务器都试图访问存储设备的情形，容易导致网络拥塞，使可靠性、安全性和稳定性变差。因此，这种网络存储比较适合小型企业，不适合数据吞吐量快速增大、并发用户数量快速增多的企业。

4.4.2.2　网络附加存储

网络附加存储（NAS）是一种采用直接与网络介质相连的特殊设备来实现数据存储的机制。由于这些设备都分配有IP地址，所以客户机通过充当数据网关的服务器可以对其进行存取访问。从结构上讲，NAS是功能单一的精简型服务器，其结构如图4.21所示。

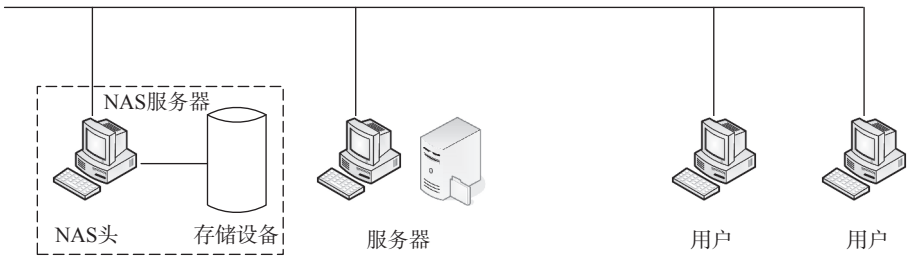


图4.21　NAS应用架构示意图

NAS是一种专业的网络文件存储及文件备份设备。它基于局域网，按照TCP/IP协议进行通信，以文件I/O方式进行数据传输。在局域网环境下，NAS已经完全可以实现异构平台之间的数据级共享，比如NT、UNIX等平台的共享。

NAS可以应用在任何网络环境中。无论是主服务器还是客户端，都可以非常方便地在NAS上存取任意格式的文件，包括SMB格式（Windows）、NFS格式（Unix，Linux）和CIFS（Common Internet File System）格式等。

NAS服务器的硬件主要包括NAS头和存储设备两部分。其中，NAS 头主要负责响应文件请求，NAS头的一端通过以太网接口连接到前端局域网，另一端通过ATA、SATA或FC等存储设备接口与后端存储设备相连（见图4.22）。存储设备通常由高性能的专有RAID磁盘阵列组成，大多 NAS 存储设备都通过对 RAID 的支持来实现数据保护和冗余，例如RAID0、RAID1、RAID5和RAID6等。

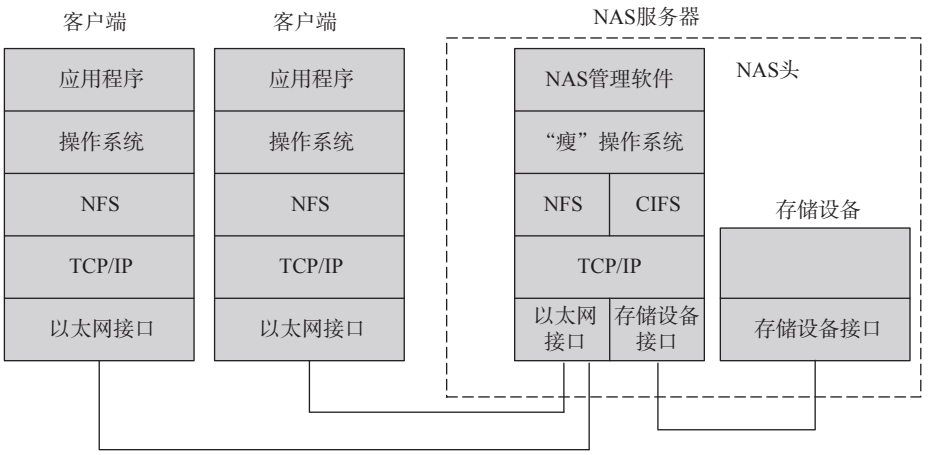


图4.22 NAS的硬件构成

NAS的软件构成主要包括以下几个主要部分。

（1）NAS操作系统

NAS 服务器上装有的操作系统是通用操作系统的裁剪版，通常称之为“瘦”操作系统，它只保留了操作系统的文件服务功能和相关文件通信协议，负责管理NAS设备的硬件和软件资源。NAS操作系统种类繁多，主要有各种Windows版、FreeBSD/Linux 版、VXWork 版以及实时操作系统RTOS等。

（2）网络文件系统

NAS上配有的网络文件系统建立在TCP/IP网络协议基础之上，用于提供远程文件访问服务，主要包括NFS和 CIFS两种。

NFS（Network File System）一般称为网络文件系统。最早由Sun公司提出，其目的是让网络上具有不同硬件、不同操作系统的各种机器可以访问彼此的文档。NFS 主要用于类 UNIX 系统中，其版本有 NFSv2（rfc1094）、NFSv3（rfc1813）和NFSv4（rfc3530）等。

CIFS（Common Internet File System）是一种主要应用于Windows系统的网络文件系统，支持网络内的计算机共享文件、打印机、串行端口和通信等资源。SMB协议是CIFS的核心，全称为服务信息块（Server Message Block），最初由IBM 在20世纪80 年代中期研发，用于将DOS的本地文件接口INT 21H转换为支持网络的文件系统。之后，微软对SMB做过多次改进，加入了许多新功能，并于1996 年将其改名为CIFS。

4.4.2.3 存储区域网络

存储区域网络（SAN）是一种通过网络方式连接存储设备和应用服务器的存储构架，专用于主机和存储设备之间的访问。当有数据存取需求时，数据可以通过存储区域网络在服务器和后台存储设备之间进行高速传输。

SAN由服务器、后端存储系统和SAN连接设备组成。后端存储系统由SAN控制器和磁盘系统构成，其中控制器是关键部分，它提供存储接入、数据操作与备份、数据共享、数据快照以及系统管理等一系列功能。使用磁盘阵列和冗余策略，可以为数据提供存储空间和安全保护措施。网络互联设备包括交换机、HBA卡和各种介质的连接线。

由于SAN是为了在服务器和存储设备之间传输大块数据而进行优化的，因此具有以下各种特点。

（1）可预计的响应时间、高可用性和可扩展性，适合关键任务数据库应用。

（2）高性能、数据一致性和可靠性，可以确保企业关键数据的安全适合集中的存储备份。

（3）可扩展的存储虚拟化，可使存储与直接主机连接相分离，并确保动态存储分区。

（4）改进的灾难容错特性，通过在主机服务器及其连接设备之间提供光纤通道，提升存储系统的性能和长距离的扩展特性。

因为SAN解决方案是从基本功能剥离出的存储功能，所以运行备份操作就必须考虑它们对网络总体性能的影响。SAN方案也使得全部存储设备都汇集在一起时的管理与集中控制相对简化。其主要用于存储数据量较大的工作环境，如ISP（信息服务供应商）、IDC（网络数据公司）、银行等。

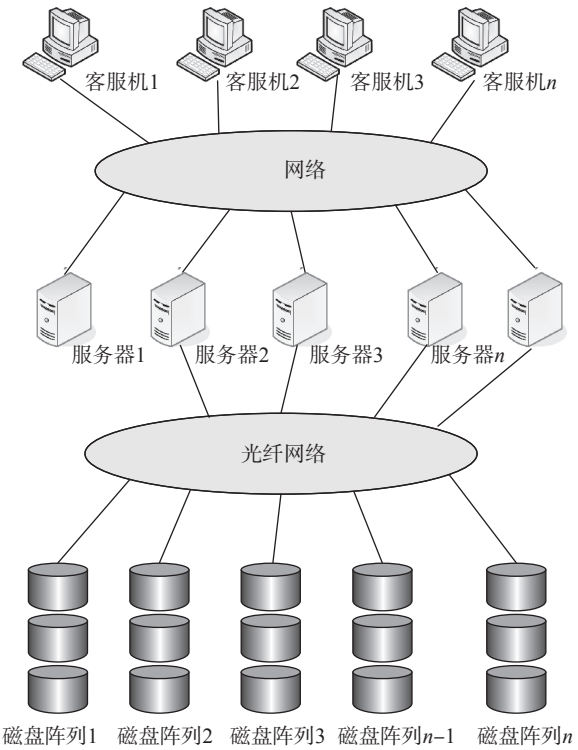


图4.23　SAN结构

4.5　容灾技术

容灾技术主要用于应付突发性灾难，如火灾、洪水、地震或者恐怖袭击等，以避免其对整个组织机构的数据和业务造成重大影响。由于企业业务越来越依赖于数据和网络服务，所以如何保证在灾难发生时关键数据不丢失，保证系统服务尽快恢复运行，成为人们日益关注的话题，容灾技术也逐渐成为人们关注的焦点。

4.5.1　容灾的分类

计算机网络系统的容灾技术涉及的内容较为广泛，与之相关的实施技术、应用和部署方案也很多。从距离的角度，可以把容灾系统分为本地容灾与异地容灾。这两种容灾类型与应用系统和备份系统之间的距离有关，容灾能力也不相同。

4.5.1.1　本地容灾

本地容灾的备份系统和业务系统部署在同一地点，可对系统的业务数据进行保护，如数据备份技术、磁盘阵列技术和数据快照技术等；也可以对应用系统的服务进行保护，如服务器集群技术和双机热备份技术等。

当由于灾难等原因导致主系统出现故障时，服务可以快速切换到备份的冗余系统，从而实现业务的快速恢复。但当灾难发生的范围比较广时，例如强烈地震或较大洪灾，可能同时损坏本地业务系统和备份的冗余系统，这时候单纯地依靠本地容灾将不能保证业务系统的安全性。

4.5.1.2　异地容灾

异地容灾是将业务系统的应用或业务数据在异地进行备份。当灾难发生后，在主系统处于灾难恢复期间时，备份系统可以接替主系统继续运行，同时对主系统的服务与数据进行恢复。

根据容灾的投资情况，也可以对原业务系统的服务进行恢复，并将备份的数据从异地恢复到本地，然后再开始业务系统的运行。一般在恢复阶段，关键业务服务的暂时停顿是不可避免的。

异地容灾系统一般由应用系统、可接替运行的异地备用系统、数据复制系统和通信线路等部分组成。为了保障系统在大灾难后的可用性，备份系统与原应用系统在距离上要足够远，网络的带宽也必须能够保证两地之间数据的顺畅同步。

异地容灾可对系统的服务、网络和数据进行保护，针对系统服务的保护包括系统失效检测和服务迁移等技术；针对数据的保护包括服务器逻辑卷备份、远程磁带备份、文件系统和数据库复制等技术。

4.5.2 容灾等级

一个容灾备份系统，需要考虑多方面的因素，如备份/恢复数据量大小、应用数据中心和备援数据中心之间的距离和数据传输方式、灾难发生时所要求的恢复速度、备援中心的管理及投入资金等。根据这些因素和不同的应用场合，通常可将容灾备份分为四个等级。

1. 本地容灾

本地容灾将系统数据或应用在本地备份，无异地后援。这一级别的容灾，仅能应付本地的硬件损坏或人为因素造成的灾难。

2. 异地数据冷备份

异地数据冷备份将系统数据备份到存储介质（磁盘、磁带或光盘）上，然后送到异地进行保存，这种方案成本低、易于实现，但是在灾难发生时，数据的丢失量大，并且系统需要很长的恢复时间，无法保持业务的连续性。

3. 异地数据热备份

异地数据热备份是指在异地建立一个热备份中心，采取同步或者异步方式，通过网络将应用系统的数据导入备份系统中。备份系统只备份数据，不承担应用系统的业务。当灾难发生时，数据丢失量小，甚至零丢失，但是应用系统恢复速度慢，无法保持业务的连续性。

4. 异地应用级容灾

异地应用级容灾是指在异地建立一个与主系统相同的备用系统，备用系统与主系统共同工作，承担系统的业务。当灾难发生时，这种容灾系统的数据丢失量很小且系统恢复速度快，能够保持业务的连续性。但是，这种方式需要配置复杂的系统管理软件和专用的硬件，成本相对来说也是最高的。

4.5.3 数据复制技术

数据复制技术是数据共享技术的一种，它将共享数据复制到多个数据存储设备中，实现了数据的本地访问，提高了数据访问效率，有效地减少了网络负荷。同时，通过同步存储设备中的数据，保证了为所有应用提供最新的相同数据。数据复制可以在多个存储设备或服务器上建立数据备份，不仅能够提高数据的安全性，提高应用程序的容灾能力，还能增强应用系统对数据的访问效率。目前，主要的数据复制技术有基于服务器逻辑卷的数据复制技术、基于存储设备的磁盘数据复制技术、基于数据库的数据复制技术和基于应用的数据复制技术。

本节主要介绍基于存储设备的磁盘数据复制技术和基于数据库的数据复制技术。

4.5.3.1　基于存储设备的磁盘数据复制技术

基于存储设备的数据复制是指依靠智能存储系统来实现数据的远程复制和同步，与服务器的操作系统和应用程序无关。也就是说，智能存储系统将服务器对存储设备执行的所有I/O操作存入日志LOG中，然后将LOG复制到远程的存储设备中，由远程的存储系统按照LOG执行I/O操作，保证数据的一致性。在这种技术下，数据复制软件运行在存储设备自带的存储系统内，与应用服务器分离，较容易实现主服务器和备用服务器的系统库、操作系统、目录和数据库的实时副本维护能力，一般不会影响主服务器系统的性能。这种数据复制技术在应用中可分为同步复制和异步复制两种方式。

1. 同步数据复制

同步数据复制是指共享数据在任何时刻、在多个复制节点间均保持一致，即主存储同步数据复制设备在备份存储设备返回操作完成的确认消息后，才给应用系统返回操作完成的确认消息。这种方式能实时保持主、备存储设备的数据一致，进而保证在灾难发生时系统能在最短时间内恢复业务运行。但是，同步数据复制加大了主服务器的工作负载，对网络速度和可靠性的要求较高，对应用系统有明显的影响，同时还要求系统能够承受由同步复制导致的时间延迟所引起的性能损失。

同步数据复制步骤如下所示。

（1）主服务器向主存储设备发出I/O操作请求；

（2）主存储设备将I/O操作请求写入存储设备控制器的缓存中，执行I/O操作，同时将I/O操作请求发送给远程备用存储设备；

（3）备用存储设备接受I/O操作请求并完成I/O操作后，将I/O操作的标识消息及操作成功消息发送到主存储设备；

（4）主存储设备在接收到备用存储设备的消息后，确认I/O操作成功，向主服务器返回操作执行成功消息。

2. 异步数据复制

异步数据复制与同步数据复制最大的不同在于，异步数据复制中的数据不是实时同步的。在异步数据复制中，主存储设备在I/O操作完成后直接返回成功消息，不会等待备用存储设备返回消息。异步数据复制中，主存储设备按照规定或需求每隔一段时间将数据同步到备用存储设备上。主要步骤如下所示。

（1）主服务器向主存储设备发送I/O操作请求；

（2）主存储设备执行数据的I/O操作，并将该I/O操作请求写进 LOG中，操作完成后向服务器返回I/O操作完成的确认消息；

（3）每隔一段时间，主存储设备将 LOG发送给远程备用存储设备；

（4）备用存储设备按照 LOG中的I/O操作请求，顺序执行I/O操作，操作完成后，将I/O操作成功消息返回给主存储设备。

同步复制技术可以确保数据的一致性和应用的完整性，而且实现简单，但却增加了应用系统的响应时间和网络负载。异步复制技术减少了应用系统的响应时间，但是实现比较复杂，且不能实时同步数据，当灾难发生时无法保证数据的完整性与一致性。

4.5.3.2 基于数据库的数据复制技术

基于数据库的数据复制是指使用数据库系统自带的软件来实现数据库的远程复制和同步，该技术的复制方式可分为同步复制和异步复制两种。基于数据库的数据复制实质是实现主服务器、备用服务器的数据库的数据同步，也就是将主服务器数据库的操作 LOG实时或者周期性地复制到备用服务器的数据库中执行。在数据复制过程中，由于数据库系统自带的软件能够自动检测和解决冲突，从而确保了数据的一致性。

(1) 同步复制是指主服务器对数据库执行的任何一个操作，都会发送到备用服务器的数据库上，备用服务器完成操作后向主服务器的数据库发送确认消息；当主服务器的数据库接收到确认消息后，向应用程序返回确认消息，然后再执行下一个操作。此种复制方式对网络可靠性的要求较高。

(2) 异步复制是指当主服务器对数据库执行操作时，数据库将这些操作按照先后顺序存储在LOG中。操作完成后会返回确认消息，不会等待备用服务器，然后每隔一段时间或按照一定的需要将LOG发送到备用服务器的数据库上。数据库按照顺序执行操作，操作完成后，发送确认消息给主服务器上的数据库，通知数据同步完成。

基于数据库的数据复制技术对主服务器的性能有一定影响，可能增加对磁盘存储容量的需求，包括对 LOG的存储，但由于系统恢复较简单，在同步复制方式时数据一致性较好。所以，对于一些数据修改更新较频繁，并且对数据一致性要求较高的应用，可采用基于数据库的远程数据复制技术。

思考题

1. 试总结比较RAID0、RAID1、RAID3、RAID5各自的存储特点，以及校验方式的特点。
2. 简述数据模型的概念和作用，以及数据模型的三要素。
3. 假设某部队的信息化中心要搭建网络存储系统，要求如下：
   (1) 该系统具有本地数据容灾的功能；
   (2) 尽量缩减搭建成本；
   (3) 绘制网络连接示意图；
   (4) 明确磁盘冗余阵列的组织方案。
   试用所学知识阐述搭建步骤。
4. 什么是范式？常用的范式有哪些？简述常用范式之间的包含关系。
5. 将下述关系规范到第三范式。
   描述超市销售信息的关系模式为(会员编号，会员姓名，购货单号，商品编码，商品名称，单价，数量，购货日期)，其数据依赖为{会员编号→会员姓名，购货单号→(购货日期，会员编号)，商品编码→(商品名称，单价)，(商品编码，购货单号)→数量}。
6. 已知关系模式 W(C, P, S, G, T, R)，函数依赖集D = {C→P, S和C→G, T和R→C, T和P→R, T和S→R}。
   其中属性的含义是：C代表课程，P代表教师，S代表学生，G代表成绩，T代表时间，R代表教室。
   (1) 计算关系模式W的一个候选键A。
   (2) W的规范化程度最高达到第几范式？
   (3) 如果关系模式W分解为3个关系模式W1(C, P)，W2(S, C, G)，W3(S, T, R, C)，则W1、W2和W3的规范化程度分别最高达到第几范式？

第5章　信息加工技术

信息加工指的是对所采集的信息用计算机进行处理，以得到更有价值的高级信息的过程。信息加工技术不是一种技术，而是一大类技术的总称。本章将介绍四种军事上比较常用的信息加工技术：数据挖掘技术、模式识别技术、信息融合技术和信息可视化技术。

5.1　数据挖掘技术

随着信息技术的发展与普及，大量的数据与信息迅速积累，出现了“可用信息稀释”的现象，也就是说，人们需要的信息混合在大量不需要的信息中，就像有价值的矿藏分布在广袤土地的某些地方，数据越多，越难以找到所需的信息。如何从海量的数据中提取出有用的和有价值的信息，就成为信息技术研究的重要问题，数据挖掘技术应运而生。

5.1.1　基本知识

5.1.1.1　什么是数据挖掘

数据挖掘（Data Mining，DM）又称数据库中的知识发现（Knowledge Discover in Database，KDD），就是从大量的、不完全的、有噪声的、模糊的甚至随机的实际应用数据中，提取出隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。

由于用来进行挖掘分析的数据大都存储在数据库（或数据仓库）中，因此有人把数据挖掘定义为“从大型数据库中的数据里提取人们感兴趣的知识”。提取的知识可以表示为概念、规则、规律和模式等形式。

原始数据可以是结构化的，如关系型数据库中的数据；也可以是半结构化的，如文本、图形、图像数据；甚至可以是分布在网络上的异构数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。

数据挖掘的用途非常广泛，可应用于生产任务的预测与分析、生产效益的评估与分析、销售领域的预测分析、物流企业的货源预测与分析、刑事案件的线索分析、超市货品的摆放、银行的贷款预测与决策分析、服装领域的职业服装号型归档、银行或商户从客户行为识别贡献最大的客户或分析高流失率客户等。在军事上也有很多应用，例如装备失效分析、装备研制管理等。

根据不同的任务，可将数据挖掘分为四类：预测建模、关联分析、聚类分析和异常检测。

5.1.1.2 数据挖掘的一般过程

数据挖掘的一般过程如图5.1所示，其中各步骤是按一定顺序完成的。当然，整个过程中还会存在步骤之间的反馈。数据挖掘的过程并不是自动的，绝大多数工作需要人工完成。在整个数据挖掘过程中，60%的时间用在数据准备上，而后续挖掘工作仅占总工作量的10%。

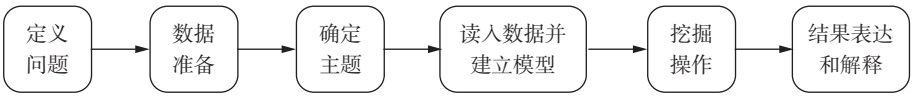


图5.1 数据挖掘的一般过程

1. 定义问题

首先明确地定义将要解决的问题。数据挖掘者要熟悉所研究行业的数据和业务问题，如果缺乏这些，就不能充分发挥数据挖掘的价值，很难得到正确的结果。

2. 数据准备

这一阶段又可分为三个子步骤：数据集成、数据选择和数据预处理。数据集成主要是将多文件或多数据库运行环境中的数据进行合并处理，解决语义模糊性，处理数据中的遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合，缩小处理范围，提高数据挖掘的质量。数据预处理是为了克服目前数据挖掘工具的局限性，提高数据质量，同时将数据转换成一个适用于特定挖掘算法的分析模型。

3. 确定主题

选择待完成的研究主题，确定待研究的合适的数据元素，以及决定如何进行数据操作等。

4. 读入数据并建立模型

用数据挖掘工具读入数据并构造出一个模型。选用的数据挖掘工具不同，构造出的数据模型也会有很大的差别。

5. 挖掘操作

利用数据挖掘工具在数据中查找。这个搜索过程可以由系统自动执行，也可以加入用户交互过程。数据挖掘的搜索过程需要反复多次，通过评价数据挖掘的结果，不断调整数据挖掘的精度，以达到发现知识的目的。

6. 结果表达和解释

根据最终用户的决策目标对提取出的信息进行分析，把最有价值的信息区分出来，并通过决策支持工具提交给决策者。

数据挖掘过程是一个多种专业人员相互配合的工作过程，不同的阶段需要具有不同专长的人员，主要包括业务分析人员、数据分析人员和数据管理人员。业务分析人员要求精通业务，能够解释业务对象，并能根据各业务对象确定用于数据定义和挖掘算法的业务需求。数据分析人员要求精通数据分析技术，对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘的各操作步骤，并为每步操作选择合适的技术。数据管理人员要求精通数据管理技术，并能从数据库或数据仓库中搜集数据。

5.1.1.3　数据挖掘系统的发展

数据挖掘软件发展到现在已经有许多种了。根据以前的一些划分依据，数据挖掘软件的发展主要经历了四代。

第一代数据挖掘软件功能比较简单，只支持一个或少数几个数据挖掘算法，并且都是一次性把数据调进内存处理，不适应大容量数据的操作。

第二代数据挖掘软件系统与数据库管理系统集成，支持数据库和数据仓库，并与它们有高性能的接口，具有较高的查询能力。

第三代数据挖掘软件系统的突出特点是能够与语言模型系统实现无缝集成，这使得由数据挖掘软件产生的模型的变化能够及时反映到语言模型系统中，从而与语言模型相联合，提供决策支持的功能。它能够挖掘网络环境下的分布式和高度异质的数据，其缺点是不支持移动环境。

第四代数据挖掘软件系统侧重于将数据挖掘和移动计算相结合，主要面向挖掘嵌入式系统、移动系统以及由常用计算设备产生的各种类型的数据。

5.1.2　预测模型

预测方法有四类典型的数学模型：(1) 回归分析，包括一元线性回归、多元线性回归、非线性回归等；(2) 趋势外推预测，包括波尔、龚珀兹、林德诺等模型；(3) 时间序列预测模型，包括移动平均、指数平滑、季节指数等；(4) 马尔可夫预测模型。这里主要介绍回归分析。

5.1.2.1　一元线性回归分析

一元线性回归分析是处理两个变量x（自变量）和y（因变量）之间关系的最简单模型，研究的是这两个变量之间的线性相关关系。这种关系可以表示如下：

$$y_i = a + bx_i + u_i$$

这个公式称为一元线性回归模型。其中，u是一个随机变量，称为随机项；a和b是两个常数，称为回归系数；i表示变量的第i个观察值，共有n组样本观察值。建立模型分成两个步骤：参数的最小二乘估计和相关性检验。

1. 参数的最小二乘估计

参数的最小二乘估计就是依据误差平方和最小的准则，估计参数a、b的值。

用$\hat{y}$、$\hat{a}$、$\hat{b}$分别表示y、a、b的估计值，$\hat{y}_i = \hat{a} + \hat{b}\hat{x}_i$，$y_i$与$\hat{y}_i$之差称为估计误差或残差，用$l_i$表示，$l_i = y_i - \hat{y}_i$。另外定义

$$l_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \ ,\ l_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \ ,\ l_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

其中，$\bar{x}$、$\bar{y}$分别是x_i、y_i的平均值，即 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

则有

$$\hat{a} = \bar{y} - \hat{b}\bar{x},\ \hat{b} = \frac{l_{xy}}{l_{xx}}$$

2. 相关性检验

相关性检验用来验证模型和实际规律符合的程度，一般用相关系数来描述变量x和y之间线性关系的密切程度。如果相关系数接近1或−1，则说明模型符合x和y的实际关系，可以用来预测；如果相关系数远离1和−1，则说明模型和实际规律符合度不够好，用来预测存在很大误差，或者不能用来预测。相关系数一般用R表示，

$$R=\frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}$$

3. 示例

有两个变量，猜测它们之间存在线性相关关系，且经检测，得到这两个变量(y,x)的4组对应值：(3.2, 1)，(4.8, 2)，(7.1, 3)，(8.9, 4)。计算它们的相关关系和相关性。

解：

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=6,\qquad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=2.5$$

$$l_{yy}=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=18.9,\qquad l_{xx}=\sum_{i=1}^{n}x_i^2-n\bar{x}^2=5$$

$$l_{xy}=\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}=9.7$$

$$\hat{b}=\frac{l_{xy}}{l_{xx}}=1.94,\qquad \hat{a}=\bar{y}-\hat{b}\,\bar{x}=1.15$$

$$R=\frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}=0.998$$

y和x的线性关系为：$y=1.15+1.94x$

由于R为0.998，非常接近1，所以y和x具有很高的相关性。

5.1.2.2 多元线性回归分析

多元线性回归分析是一元线性回归分析的推广，其公式为

$$y_i=b_0+b_1x_{1i}+b_2x_{2i}+\cdots+b_kx_{ki}+u_i,\qquad i=1,2,\cdots,n$$

1. 参数的最小二乘估计

对应的样本回归模型为

$$\hat{y}_i=\hat{b}_0+\hat{b}_1x_{1i}+\hat{b}_2x_{2i}+\cdots+\hat{b}_kx_{ki}$$

则有

$$\hat{\boldsymbol{B}}=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y})$$

其中，$\boldsymbol{x}=\begin{bmatrix}1 & x_{11} & x_{21} & \cdots & x_{k1}\\ 1 & x_{12} & x_{22} & \cdots & x_{k2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn}\end{bmatrix}$，$\boldsymbol{y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$，$\hat{\boldsymbol{B}}=\begin{bmatrix}\hat{b}_0\\ \hat{b}_1\\ \vdots\\ \hat{b}_k\end{bmatrix}$，$\boldsymbol{x}^{\mathrm{T}}$为$\boldsymbol{x}$的转置矩阵。

2. 多元线性回归模型的检验

$\mathrm{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 表示观察值y_i与其平均值的总离差平方和。

$\mathrm{ESS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 表示由回归方程中x的变化引起的y的变化，称为回归平方和。

$\mathrm{RSS}=\mathrm{TSS}-\mathrm{ESS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$表示不能用回归方程解释的部分，是由其他未能控制的随机干扰因素引起的残差平方和。

检验包括拟合优度检验、回归方程的显著性检验（F检验）、回归系数的显著性检验（t检验）。这里只给出拟合优度检验。拟合优度：

$$R^2=\mathrm{ESS}/\mathrm{TSS},\qquad 0\leqslant R^2\leqslant 1$$

R^2越接近1，则回归直线与样本观察值拟合得越好。

5.1.2.3　非线性回归预测模型

在许多实际问题中，不少变量之间的关系是非线性的，可以通过变量代换把本来应该用非线性回归处理的问题近似转化成线性回归问题，再进行分析预测。表5.1中列举的是5种常见的非线性模型及线性变换的方式。这些非线性模型都可转化为一元或多元线性模型，可以利用前面介绍过的一元和多元线性回归模型的最小二乘法求出参数估计、模型的拟合优度等。

表5.1　5种常见的非线性模型及线性变换的方式

幂函数模型	$y=ax^b$	$y'=\lg(y)$ $x'=\lg(x)$ $a'=\lg(a)$	$y'=a'+bx'$
双曲线模型	$1/y=a+b(1/x)$	$y'=1/y$ $x'=1/x$	$y'=a+bx'$
对数函数模型	$y=a+b\lg(x)$	$x'=\lg(x)$	$y=a+bx'$
指数函数模型	$y=ae^{bx}$	$y'=\ln(y)$ $a'=\ln(a)$	$y'=a'+bx$
多项式曲线模拟	$y=b_0+b_1x+b_2x^2+\cdots+b_kx^k$	$x_1=x$ $x_2=x^2$ $\cdots$ $x_k=x^k$	$y=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k$

5.1.3　关联分析

关联分析又称关联挖掘，是基于关联规则的数据挖掘技术。关联规则（Association Rule）是指在大型的数据库系统中，迅速找出各事物之间潜在的、有价值的关联，用规则表示出来，经过推理、积累形成知识后，得出重要的相关联的结论，从而为决策提供支持。

关联规则的研究和应用是数据挖掘中最活跃和比较深入的分支之一，目前已经提出了许多关联规则挖掘的理论和算法。其中使用较多的是支持度–置信度框架。其定义如下。

设I = $\{i_1, i_2, \cdots, i_n\}$是项的集合。设任务相关的数据D是数据库事务的集合，其中每个事务T是项的集合，使得T⊆I。每个事务有一个标识符，称为事务ID。设A是一个项集，事务T包含A当且仅当A⊆T。关联规则是形如A⇒B的形式，其中A⊂I，B⊂I，并且A∩B=∅。规则A⇒B在事务D中成立，具有支持度s，其中s是D中事务包含A或B的百分比，它的概率为P(A或B)。如果D中包含A的事务同时也包含B的百分比是c，那么规则A⇒B在事务集D中具有置信度c。这是条件概率P(B|A)，即

$$s = P(\text{A或B}) = \frac{\text{包含A或B的T的数量}}{\text{D中T的总数}}$$

$$c = P(\text{B|A}) = \frac{\text{同时包含A和B的T的数量}}{\text{包含A的T的数量}}$$

支持度是项集中包含A或B的事务数与所有事务数之比，描述了A和B在所有事务中出现的概率有多大，说明了规则的有用性。规则A⇒B的置信度，是指在出现了A的事务中，事务B出现的概率有多大，说明了规则的确定性。

为了发现有意义的关联规则，需要给定两个阈值：最小支持度（Minimum Support）和最小置信度（Minimum Confidence）。挖掘出的关联规则必须满足用户规定的最小支持度，它表示了一组项目关联在一起所需满足的最低联系程度。挖掘出的关联规则也必须满足用户规定的最小置信度，它反映了一个关联规则的最低可靠程度。在这个意义上，数据挖掘系统的目的就是从数据库中挖掘出满足最小支持度和最小置信度的关联规则。

例如购物篮分析。表5.2给出的是在一家杂货店收银台收集的销售数据。关联分析可以用来发现顾客经常同时购买的商品。

设A={尿布}，B={牛奶}

$$s = P(\text{A或B}) = 5/10 = 0.5,\ c = P(\text{A}\Rightarrow\text{B}) = 5/5 = 1$$

如果设最小支持度为0.4，最小置信度为0.7，就可以认为发现了规则{尿布}→{牛奶}。该规则暗示购买尿布的顾客多半会购买牛奶。这种类型的规则可以用来发现各类商品中可能存在的交叉销售的商机。

表5.2 购物篮数据

事务ID	商品
1	{面包，黄油，尿布，牛奶}
2	{咖啡，糖，小甜饼，鲑鱼}
3	{面包，黄油，咖啡，尿布，牛奶，鸡蛋}
4	{面包，黄油，鲑鱼，鸡}
5	{鸡蛋，面包，黄油}
6	{鲑鱼，尿布,牛奶}
7	{面包，茶，糖，鸡蛋}
8	{咖啡，糖，鸡，鸡蛋}
9	{面包，尿布，牛奶，盐}
10	{茶，鸡蛋，小甜饼，尿布，牛奶}

关联规则主要适合应用于大量数据且其表面不呈现规律的情况，目的是发现新的关联关系。成功的关联挖掘就是发现了新的关联关系的挖掘。

5.1.4　聚类分析

聚类是将数据划分成群组的过程。通过确定数据之间在预先指定的属性上的相似性来完成聚类任务，这样最相似的数据就聚集成簇，也就是可以作为一类。聚类与分类不同，聚类的类别取决于数据本身，而分类的类别是由数据分析人员预先定义好的。使用聚类算法的用户不但需要深刻了解所用的特殊技术，而且还要了解数据收集过程的细节，以及拥有应用领域的专家知识。

现有的聚类技术大致可以分为五种：基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法和基于模型的方法。下面重点介绍前三种方法。

5.1.4.1　基于划分的聚类方法

给定N个对象（数据），把它们分成K组，要求每个组中的对象尽可能“相似”，即差别尽量小，不同组中对象之间的差别则尽可能大。

基于划分的聚类方法中，最经典的就是k平均（k-means）算法和k中心（k-medoids）算法，很多算法都是由这两个算法改进而来的。

> k平均算法描述如下。
> 输入：聚类个数k，以及包含n个数据对象的数据库。
> 输出：满足方差最小标准的k个聚类。
>
> 处理流程：
> Step1　从n个数据对象中任选k个对象作为初始聚类中心；
> Step2　针对每个对象，计算对象与每个聚类中心的距离，并把该对象划分到距离最小的类中，循环Step2，直到处理完所有对象；
> Step3　计算每一个类的平均值，作为该类的新的聚类中心；
> Step4　循环Step2到Step3，直到所有的类都不再发生变化为止。

k平均算法的工作过程如下：首先从n个数据对象中任选k个对象作为初始聚类中心；对于其余对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的聚类中心所代表的类；然后再计算每个新类的聚类中心（该类中所有对象的平均值）；不断重复这一过程，直到标准测度函数收敛。

一般采用均方差作为标准测度函数，即准则函数。两个数据对象，如x_i和y_i之间的距离有多种定义，常见的包括明氏距离、欧氏距离、马氏距离、兰氏距离等。其中欧氏距离就是我们常用的距离定义，即

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^2}$$

其中，p代表数据对象的维数，k代表数据的第k个属性。

准则函数可以表示为

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, z_i)$$

其中，C_i代表第i类的样本集合。也就是说，先求出每个类的平方误差和，再把所有的平方误差和求总和。E越小，说明分类效果越好。

例 设有数据样本集合为X = {1, 5, 10, 9, 26, 32, 16, 21, 14}，将X聚为3类，即$k = 3$。相似度采用欧氏距离计算。迭代过程如下所示。

第一次迭代：随机选择前三个数值为初始的聚类中心，即$z_1 = 1$，$z_2 = 5$，$z_3 = 10$。按照三个聚类中心将样本集合分为如下三类：

$$C_1=\{1\},\quad C_2=\{5\},\quad C_3=\{10, 9, 26, 32, 16, 21, 14\}$$

第二次迭代：分别计算三个类的平均值，作为新的聚类中心，$z_1 = 1$，$z_2 = 5$，$z_3 = 18.3$。按照新的聚类中心重新聚类。结果为

$$C_1=\{1\},\quad C_2=\{5, 10, 9\},\quad C_3=\{26, 32, 16, 21, 14\}$$

第三次迭代：分别计算三个类的平均值，作为新的聚类中心，$z_1 = 1$，$z_2 = 8$，$z_3 = 21.8$。按照新的聚类中心重新聚类，结果为

$$C_1=\{1\},\quad C_2=\{5, 10, 9, 14\},\quad C_3=\{26, 32, 16, 21\}$$

第四次迭代：分别计算三个类的平均值，作为新的聚类中心，$z_1 = 1$，$z_2 = 9.5$，$z_3 = 23.8$。按照新的聚类中心重新聚类，结果为

$$C_1=\{1, 5\},\quad C_2=\{10, 9, 14, 16\},\quad C_3=\{26, 32, 21\}$$

第五次迭代：分别计算三个类的平均值，作为新的聚类中心，$z_1 = 3$，$z_2 = 12.3$，$z_3 = 26.3$。按照新的聚类中心重新聚类，结果为

$$C_1=\{1, 5\},\quad C_2=\{10, 9, 14, 16\},\quad C_3=\{26, 32, 21\}$$

迭代结果不变，准则函数E收敛，迭代结束。如表5.3所示。

表5.3 *k*–means聚类算法

步骤	z_1	z_2	z_3	C_1	C_2	C_3	E
1	1	5	10	1	5	10, 9, 26, 32, 16, 21, 14	433.43
2	1	5	18.3	1	5, 10, 9	26, 32, 16, 21, 14	230.8
3	1	8	21.8	1	5, 10, 9, 14	26, 32, 16, 21	181, 76
4	1	9.5	23.8	1, 5	10, 9, 14, 16	26, 32, 21	101.43
5	3	12.3	26.3	1, 5	10, 9, 14, 16	26, 32, 21	101.43

*k*中心算法描述如下。

输入：聚类个数*k*，以及包含*n*个数据对象的数据库。

输出：满足方差最小标准的*k*个聚类。

处理流程：

Step1 从*n*个数据对象任选*k*个对象作为初始类的中心点；

Step2 将每个剩余的对象指派给离它最近的中心点所代表的类；

Step3 选择一个未被选择的中心点对象O_i；

Step4 选择一个未被选择过的非中心点对象O_h；

Step5 计算用O_h替代O_i的总代价，并记录在集合S中。

Step6 循环Step4到Step5，直到所有的非中心点都被选择过；

Step7 循环Step3到Step6，直到所有的中心点都被选择过；

Step8 IF在S中的所有非中心点代替所有中心点后计算出的总代价有小于0的存在，THEN找出S的中心点，形成一个新的*k*个中心点的集合；

Step9 循环Step3到Step8，直到没有再发生类的重新分配，即S中所有的元素都大于0。

k中心聚类方法的基本思路是：选用类中位置最中心的对象作为代表对象，试图对n个对象给出k个划分。代表对象又称为中心点，其他对象则称为非代表对象。最初随机选择k个对象作为中心点，然后反复用非代表对象来代替代表对象，试图找出更好的中心点，以改进聚类的质量。在每次迭代中，要分析所有可能的对象对，每个对中的一个对象是中心点，而另一个是非代表对象。对可能的各种组合，估算聚类结果的质量。一个对象O_i可以被使最大平方误差值E最小的对象代替。在一次迭代中产生的最佳对象集合成为下一次迭代的中心点。

为了判定一个非代表对象O_h是否可以代替当前的代表对象O_i，对于每一个非中心点对象O_j，考虑如下四种情况：

(1) O_h代替O_i作为新的中心点，O_j当前隶属于O_i。如果O_j离某个中心点O_m（$m \neq i$）最近，那么O_j重新分配给O_m。

(2) O_h代替O_i作为新的中心点，O_j当前隶属于O_i。如果O_j离新的中心点O_h最近，则O_j重新分配给O_h。

(3) O_h代替O_i作为新的中心点，O_j当前隶属于另一个中心点O_m。如果O_j离O_m最近，则对象隶属关系不变。

(4) O_h代替O_i作为新的中心点，O_j当前隶属于另一个中心点O_m。如果O_j离O_h最近，则O_j重新分配给O_h。

每当发生重新分配时，E所产生的差别对代价函数会有影响。因此，如果一个当前的中心点对象被非中心点对象所代替，则代价函数计算E所产生的差别。替换的总代价是所有非中心点对象所产生的代价之和。如果总代价是负的，那么实际的E将会减少，O_h可以代替O_i。如果总代价是正的，则当前中心点O_i不变。总代价定义如下：

$$\mathrm{TC}_{ih} = \sum_{j=1}^{n} C_{jih}$$

其中，C_{jih}表示O_h代替O_i后产生的代价。

设空间中有五个点$\{A, B, C, D, E\}$，各点之间的距离关系如表5.4所示。根据所给的数据对其运行k-medoids算法，实现聚类划分（$k = 2$）。

表5.4　样本点之间的距离

样本点	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

算法执行步骤如下。

第一步：设从5个对象中随机抽取的2个中心点为A和B，则样本被划分为$\{A, C, D\}$和$\{B, E\}$（与两个中心点距离相同的非中心点随机划分到两个类）。

第二步：假定中心点A和B分别被非中心点C, D和E替换。根据算法，需要计算下列代价TC_{AC}，TC_{AD}，TC_{AE}，TC_{BC}，TC_{BD}，TC_{BE}。其中，TC_{AC}表示中心点A被非中心点C代替后的总代价。下面以TC_{AC}为例说明计算过程。

当A被C替换以后，下面列出了各对象的变化情况。

(1) A：A不再是一个中心点，C称为新的中心点，因为A离B比A离C近，A被分配到B中心点所代表的类，属于前述第一种情况。$C_{AAC} = d(A, B) - d(A, A) = 1 - 0 = 1$。

（2）B：B不受影响，属于上面的第三种情况。$C_{BAC}=0$。

（3）C：C原先属于A中心点所在的类，当A被C替换以后，C是新中心点，属于上面的第二种情况。$C_{CAC}=d(C,C)-d(A,C)=0-2=-2$。

（4）D：D原先属于A中心点所在的类，当A被C替换以后，离D最近的中心点是C，属于上面的第二种情况。$C_{DAC}=d(D,C)-d(D,A)=1-2=-1$。

（5）E：E原先属于B中心点所在的类，当A被C替换以后，离E最近的中心点仍然是B，属于上面的第三种情况。$C_{EAC}=0$。

因此，$TC_{AC}=C_{AAC}+C_{BAC}+C_{CAC}+C_{DAC}+C_{EAC}=1+0-2-1+0=-2$。同理，可以计算出$TC_{AD}=-2$，$TC_{AE}=-1$，$TC_{BC}=-2$，$TC_{BD}=-2$，$TC_{BE}=-2$。在上述代价计算完毕后，选取一个最小的代价，显然有多种选择，选择第一个最小代价的替换，即C替换A。这样，样本被重新划分为$\{A,B,E\}$和$\{C,D\}$两个类。通过上述计算，已经完成了第一次迭代。在下一次迭代中，将分别用其他的非中心点A, D, E替换中心点B和C，找出具有最小代价的替换。一直重复上述过程，直到代价不再减少为止。

5.1.4.2 基于层次的聚类方法

基于层次的聚类方法对给定的数据进行层次分解，直到满足某种条件为止。首先将数据对象组成一棵聚类树，然后根据层次，自底向上或自顶向下进行分解。层次的方法可以分为凝聚的方法和分裂的方法。

1. *凝聚的方法*

凝聚的方法又称为自底向上的方法，初始时将每个对象都看成单独的一个类，然后通过逐步地合并相近的类而形成越来越大的类，直到所有对象都在一个类中，或者满足某个终止条件为止。层次凝聚的代表是AGNES（AGglomerative NEString）算法。

AGNES算法描述如下。
输入：包含n个数据对象的数据库，终止时类的数目为k。
输出：满足终止条件规定的k个类。

处理流程：
Step1　将每个对象当成一个初始类；
Step2　根据两个类中最近的数据点找到距离最近的两个类；
Step3　合并两个类，生成新的类；
Step4　循环Step3到Step4，直到类的数目达到k。

例　有一组数据：{1, 5, 7, 12, 13, 15, 20, 22}，以数据差值的绝对值作为数据之间的距离，用AGNES算法对这组数据聚类。

用{}表示数据类别。根据AGNES算法，首先把每个数据作为一个初始类，如下所示：

{1} {5} {7} {12} {13} {15} {20} {23}

然后，选择其中距离最近的两个类合并，显然{12}和{13}两个类的距离最近，所以把它们合并成{12, 13}，作为新类，聚类结果如下：

{1} {5} {7} {12, 13} {15} {20} {23}

依次类推，不断合并距离最近的类，结果如下：

{1}　{5, 7}　{12, 13}　{15}　{20}　{23}
{1}　{5, 7}　{12, 13, 15}　{20}　{23}
{1}　{5, 7}　{12, 13, 15}　{20, 23}
{1, 5, 7}　{12, 13, 15}　{20, 23}
{1, 5, 7, 12, 13, 15}　{20, 23}
{1, 5, 7, 12, 13, 15, 20, 23}

如果k为1，则最终把所有的数据合并成一个类别。如果$k > 1$，则在聚类数量等于k时，停止聚类，当前的聚类状态就是结果。

2. 分裂的方法

分裂的方法又称为自顶向下的方法，初始时将所有的对象作为一个类，然后逐渐细分为更小的类，直到最终每个对象都在单独的一个类中，或者满足某个终止条件为止。层次分裂的代表是DIANA（DIvisive ANAlysis）算法。DIANA算法中使用了下面两种测度方法。

（1）类的直径。在一个类中的任意两个数据点都有一个距离（如欧氏距离），这些距离中的最大值是类的直径。

（2）平均相异度（平均距离）。表示为

$$d_{\text{avg}}(x,C)=\frac{1}{n-1}\sum_{y\in C,y\neq x}d(x,y)$$

其中，$d_{\text{avg}}(x, C)$表示点x在类C中的平均相异度，n为类C中点的个数，$d(x, y)$为点x与点y之间的距离（如欧氏距离）。

DIANA算法描述如下。
输入：包含n个数据对象的数据库，终止时类的数目k。
输出：满足终止条件规定的k个类。

处理流程：
Step1　将所有对象整个当成一个初始类；
Step2　在所有类中挑选出具有最大直径的类；
Step3　找出所挑选类里与其他点平均相异度最大的一个点放入splinter group，剩余的放入old party中；
Step4　在old party中找出到splinter group中点的最近距离不大于到old party中点的最近距离的点，并将该点加入splinter group；
Step5　循环Step2到Step4直到没有新的old party的点分配给splinter group；
Step6　splinter group和old party为被选中的类分裂成的两个类，与其他类一起组成新的类集合；
Step7　循环Step2到Step6，直到类的数目达到k。

例　有一组数据{1, 5, 7, 12, 13, 15, 20, 22}，用DIANA算法对这组数据聚类。

根据DIANA算法，首先把所有的数据作为一个类别。通过计算可知，该类的直径为21，数据1与其他数据的相异度最大，为12.4。找到直径最大的类，找出平均相异度最大的数据。为了看起来方便，把该类直径、该数据及其相异度写在该类的后面，如下所示：

{1, 5, 7, 12, 13, 15, 20, 22} 21 1 12.4

把1取出来，放入Splinter group，其余数据作为old party，如下所示：

Splinter group {1} old party {5, 7, 12, 13, 15, 20, 22}

从old party中，查找和Splinter group中数据的最近距离不大于和old party中数据的最近距离的点，如果有，将该数据加入Splinter group中，循环处理old party中每一个数据。处理完毕，把Splinter group和old party作为新类。结果如下：

{1} {5, 7, 12, 13, 15, 20, 22}

同样，下面依次处理。

{1} {5, 7, 12, 13, 15, 20, 22} 17 22 10
{1} old party {5, 7, 12, 13, 15, 20} Splinter group {22}
{1} old party {5, 7, 12, 13, 15} Splinter group {20, 22}
{1} {5, 7, 12, 13, 15} {20, 22}
{1} {5, 7, 12, 13, 15} 10 5 6.75 {20, 22}
{1} Splinter group {5} old party {7, 12, 13, 15} {20, 22}
{1} Splinter group {5, 7} old party {12, 13, 15} {20, 22}
{1} {5, 7} {12, 13, 15} {20, 22}
{1} {5, 7} {12, 13, 15} 3 15 2.5 {20, 22}
{1} {5, 7} old party {12, 13} Splinter group {15} {20, 22}
{1} {5, 7} {12, 13} {15} {20, 22}
{1} {5, 7} 2 5 2 {12, 13} {15} {20, 22}
{1} Splinter group {5} old party {7} {12, 13} {15} {20, 22}
{1} {5} {7} {12, 13} {15} {20, 22}
{1} {5} {7} {12, 13} {15} {20, 22} 2 20 2
{1} {5} {7} {12, 13} {15} Splinter group {20} old party {22}
{1} {5} {7} {12, 13} {15} {20} {22}
{1} {5} {7} {12 13} 1 12 1 {15} {20} {22}
{1} {5} {7} Splinter group {12} old party {13} {15} {20} {22}
{1} {5} {7} {12} {13} {15} {20} {22}

如果$k = n$，则把每个数据分为一类，算法终止。如果$k < n$，则当分类的个数等于k时，算法终止。

5.1.4.3 基于密度的聚类方法

基于密度的方法与其他方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的，这样就能克服基于距离的算法只能发现球状聚类，对发现任意形状的聚类则显得不足的缺点。基于密度的聚类方法从对象分布区域的密度着手，对于给定类中的数据点，如果在给定范围的区域内的对象或数据点的密度超过某一阈值，就继续聚类。这样，通过连接密度较大的区域就能形成不同形状的聚类，而且还可以消除孤立点和噪声对聚类质量的影响，发现任意形状的簇。

下面是关于密度聚类涉及的一些定义。

（1）对象的ε邻域：给定对象在半径ε内的区域。

（2）核心对象：如果一个对象的ε邻域至少包含MinPts个对象，则称该对象为核心对象。

（3）直接密度可达：给定一个对象集合D，如果p是在q的ε邻域内，而q是一个核心对象，则对象p从对象q出发是直接密度可达的。

（4）间接密度可达：如果存在一个对象链 $p_1, p_2, \cdots, p_n$，$p_1 = q$，$p_n = p$，对$p_i \in$D，$1 \leqslant i \leqslant n$，$p_{i+1}$是从$p_i$关于$\varepsilon$和MinPts直接密度可达的，则对象$p$是从对象$q$关于$\varepsilon$和MinPts间接密度可达的。

（5）密度相连：如果对象集合D中存在一个对象o，使得对象p和q是从o关于ε和MinPts密度可达的，那么对象p和q是关于ε和MinPts密度相连的。

（6）噪声：一个基于密度的类是基于密度可达性的最大的密度相连对象的集合。不包含于任何类中的对象被认为是“噪声”。

DBSCAN算法描述如下。

输入：包含n个数据对象的数据库，半径ε，最少数目MinPts。

输出：所有达到密度要求的类。

处理流程：

Step1　从数据库中抽取一个未处理的点；

Step2　IF 抽出的点是核心点THEN找出所有从该点密度可达的对象，形成一个类；

Step3　ELSE 抽出的点是边缘点（非核心对象），跳出本次循环，寻找下一个点；

Step4　循环Step1到Step3直到所有的点都已处理。

例　根据文章的主题词进行文章聚类，相同的主题词越多，则文章越相似，距离越近。表5.5收集了8篇文章的主题词。我们对这些文章进行聚类。

表5.5　新闻文章集合

文章	词
1	dollar，industry，country，loan，deal，director
2	market，industry，work，country
3	dollar，market，country，index
4	country，market，sale，industry，price
5	patient，symptom，drug，health，clinic，doctor
6	health，company，drug，patient
7	drug，public，health，medical，director
8	medical，cost，patient，health，case，director

为了对文章进行聚类，首先定义文章之间的距离为

$$d(x, y) = \frac{c_d}{c_s + c_d}$$

其中，$d(x, y)$表示文章x和文章y之间的距离。c_s表示两篇文章之间相同主题词的数量，c_d表示两篇文章之间不同主题词的数量。

很显然，根据定义，同一篇文章自己和自己的距离为0；对于文章x和y，$d(x, y) = d(y, x)$。设每篇文章的主题词数量相等，则两篇文章相同词越多时距离越近；不同词越多时距离越远。如果每篇文章主题词数量不同，则差距越大时距离越远。符合我们的一般认识。

例如，文章1和文章2有两个相同词：industry和country，其他词不同，那么它们之间的距离为

$$d(1,2)=\frac{6}{2+6}=0.75$$

类似地，可以计算出任何两篇文章之间的距离，针对表5.5的8篇文章，计算出它们之间的距离如表5.6所示。

表5.6 基于主题词定义的文章之间的距离

	文章1	文章2	文章3	文章4	文章5	文章6	文章7	文章8
文章1	0	0.75	0.75	0.667	1	1	0.9	0.917
文章2	0.75	0	0.667	0.5	1	1	1	1
文章3	0.75	0.667	0	0.714	1	1	1	1
文章4	0.667	0.5	0.714	0	1	1	1	1
文章5	1	1	1	1	0	0.571	0.778	0.8
文章6	1	1	1	1	0.571	0	0.714	0.8
文章7	0.9	1	1	1	0.778	0.714	0	0.625
文章8	0.917	1	1	1	0.8	0.8	0.625	0

如果我们把ε设为0.85，MinPts设为3，则用DBSCAN算法可以把上述文章分为两类，前4篇文章为一类，后4篇文章为一类。

5.2 模式识别技术

模式识别诞生于20世纪20年代，随着40年代计算机的出现，50年代人工智能的兴起，模式识别在60年代初迅速发展成为一门学科，并获得了广泛应用。本节将介绍有关模式识别的一些基本概念和技术。

5.2.1 基本知识

5.2.1.1 什么是模式识别

我们在生活中时时刻刻都在进行模式识别，例如能认出周围的物体是桌子、椅子，能认出对面的人是张三、李四，能区分各种不同的声音、气味等。

从不同的角度看同一个人的脸，虽然视网膜上的成像不相同，但我们可以认出这个人。换句话说，不同角度的像都具有一种相同的模式。我们所关注的不是像本身，而是像所包含的模式。虽然我们每次看一个人的角度不同，但总能够认出他来，因为我们能够识别不同的像所具有的相同模式。我们也能够区分不同的人具有的不同模式，所以可以认出不同的人。如果两个人长得很像，我们可能会认错人，这就是因为无法区分两个人的像所具有的模式了。

显然，模式指的不是事物本身，而是我们从事物获得的信息在时间或空间上的分布规律或特征。

在计算机出现以后，人们试图用计算机来实现人或动物所具备的模式识别的能力，于是诞生了模式识别这门学科。提到“模式识别”一词时，主要是指计算机进行的模式识别。计算机要识别信息所具有的模式，首先就要把信息变成计算机中的信息，或者说是数字信息，然后用计算机分析信息所具有的模式。

需要注意，这里的时间和空间应按广义理解。例如，医生根据各项化验指标判断疾病种类，这也是模式识别，这时各种化验项目并不对应物理的时间或空间，但可以看成数学或逻辑上的空间。

将每种模式看成一个特征样本，类似的模式即可构成一个类别。如果给每个类别命名，并且用特定的符号来表达这个名字，那么模式识别可以看成从具有时间和空间分布的信息向着符号所进行的映射。模式识别的过程就成为一个分类的过程，即对于一个待分类的事物，把它分到已有的类别中；如果该事物和任何已有类别都不相同，则为该事物建立一个新的类别。

5.2.1.2　模式识别系统

有两种基本的模式识别方法，即统计模式识别方法和结构（句法）模式识别方法。基于这两种方法的模式识别系统都由两个过程组成，即设计和实现。设计是指用一定数量的样本（称为训练集或学习集）进行分类器的设计。实现是指用所设计的分类器对待识别的样本进行分类决策。我们只介绍统计模式识别方法。

如图5.2所示，基于统计方法的模式识别系统主要由4部分组成：数据获取、预处理、特征提取和分类决策（或特征选择和分类器设计）。其中包含了设计和实现两个过程，这两个过程的前三部分都相同，设计过程的第四部分是特征选择和分类器设计，实现过程的第四部分是分类决策。

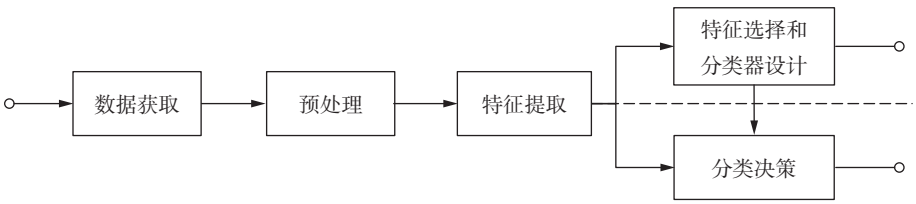


图5.2　模式识别系统的基本构成

1. 数据获取

数据获取就是把模式识别的对象用计算机可以处理的数据表示出来。例如，人脸识别中，数据获取就是通过数字摄像设备得到人脸的照片文件。获取的数据通常有下列3种类型，即

(1) 二维图像。如文字、指纹、地图、照片等。

(2) 一维波形。如脑电图、心电图、机械振动波形等。

(3) 物理参量和逻辑值。物理参量的例子有：在疾病诊断中病人的体温及各种化验数据等；对某参量正常与否的判断或对症状有无的描述，如疼与不疼，则可用逻辑值表示。在引入模糊逻辑的系统中，这些值还可以包括模糊逻辑值，比如很大、大、比较大等。

通过测量、采样和量化，可以用矩阵或向量表示二维图像或一维波形。

2. 预处理

预处理的目的是去除噪声，加强有用的信息，并对输入设备、测量仪器或其他因素造成的退化现象进行复原。

3. 特征提取

由图像或波形所获得的数据量是相当大的。例如，一个文字图像可以有几千个数据，一个心电图波形也可能有几千个数据，一个卫星遥感图像可能有几千万项甚至更多的数据。为了有

效地实现分类识别，就要对原始数据进行变换，得到最能反映分类本质的特征。这就是特征提取和选择的过程。

一般我们把原始数据组成的空间称为测量空间，把分类识别赖以进行的空间称为特征空间，通过变换，可以把在维数较高的测量空间中表示的模式变为在维数较低的特征空间中表示的模式。特征空间中的模式通常又称为样本，它往往可以表示为一个向量，即特征空间中的一个点。

4. 分类决策（特征选择和分类器设计）

这个部分在实现过程是分类决策，在设计过程则是特征选择和分类器设计。

分类决策就是在特征空间中用统计方法把被识别对象归为某一类别。基本做法是在样本训练集基础上确定某个判决规则，使得根据这种判决规则对被识别对象进行分类所造成的错误识别率最小，或引起的损失最小。

特征选择和分类器设计属于设计过程。从事物的数据中可以提取很多特征，这些特征具有不同的分类能力。对于我们想划分的类别，有些特征在各类别之间差别不大，即分类能力弱；有些特征在各类别之间差别很大，即分类能力强。我们需要的是那些分类能力强的特征，使用这些特征进行分类器的设计。

关于数据获取和预处理的理论和技术是数字信号处理和图像处理的研究课题，一般与具体研究的问题密切相关。特征提取、特征选择和分类器设计、分类决策是模式识别的主要研究内容。

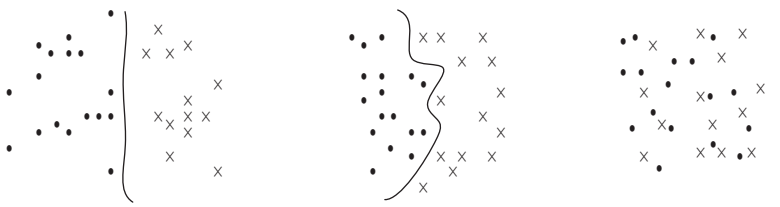
5.2.1.3 几个基本问题

关于模式识别，有几个基本问题很重要，但目前还没有令人满意的解决方法，这里进行简要说明。

1. 模式类的紧致性

为了能在某个特征空间中进行分类，通常假设同一类的各个模式在该特征空间中组成一个紧致集。从这个紧致集中的任何一点可以均匀地过渡到同一集中的另外一点，而在过渡途中的所有各点都仍然属于这个紧致集，即属于同一模式类。此外，当紧致集中的各点在任意方向有某些不大的移动（相应于被观察现象有某些微小的变形）时，它仍然属于这个集合。

如图5.3所示，考虑两个类别的3种情况，其中(a)没有临界点，(b)有许多临界点，(c)临界点多到使分类不可能实现。



(a) (b) (c)

图5.3 3种临界点情况

紧致集具有下列性质：

(1) 临界点的数量与总的点数相比很少。

(2) 集合中任意两个内点可以用光滑线连接，在该连线上的点也属于这个集合。

(3) 每个内点都有一个足够大的邻域，在该邻域中只包含同一集合中的点。

假设每个模式类都满足紧致性假设，则解决模式识别问题就不会遇到什么原则上的困难。对于很多实际问题，这个假设是不成立的。但是，如果我们能够把实际中的问题进行变换，让它满足紧致性假设，那就可以很容易地进行识别了。所以，我们换句话来说明模式识别：模式识别就是寻找一种方法，使测量空间中属于同一类的所有各点都映射到特征空间中的同一个紧致集，而把另一类的所有各点都映射到特征空间中的另外一个紧致集，且使这两个紧致集相隔一个显著的距离。显然，在这个特征空间中，模式类是满足紧致性假设的。

只要各个模式类是可分的，总存在这样一个空间，使变换到这个空间中的集合满足紧致性要求。这样一种变换只能在解决识别任务的过程中来求取，与具体问题紧密相关。

2. 相似与分类

模式识别是把具体事物归入某一类别的过程。要进行分类，首先要有类存在。通常设计模式识别系统时，分类标准是人为地从系统外给定的，通过设计或有监督的学习过程，使系统能完成特定的识别任务。这种方法的优点是能设计出较经济的系统，软件规模不会过于庞大；缺点是系统学习能力不强。

另一种方法是让系统自动对事物进行分类，把相似的事物自动归为一类。其优点是系统学习能力强；缺点是软件复杂、庞大，分类和识别结果具有一定的不确定性。由于有很多情况需要尽量减少人的干预，增加系统本身的适应能力，因此这种方法越来越得到广泛的应用。

那么系统如何进行分类呢？如何判断事物之间的相似性呢？首先从集合的有关概念来说明。

用集合论中的子集和元素来代表模式类和模式。在一个集合M中，可以定义一个关系R，如果对所有的$x\in M$，xRx成立，则称R是自反的；如果对所有x，$y\in M$，$xRy \Rightarrow yRx$成立，则称关系R是对称的。相似关系是满足对称和自反的关系。如果对于x，y，$z\in M$，xRy，$yRz \Rightarrow xRz$，则称关系R是传递的。同时满足自反、对称和传递的关系称为等价关系。例如，相等就是一种等价关系。

满足等价关系的集合必定可以划分为若干子集，即 $M=\bigcup_i M_i$ 且 $M_i \cap M_j = \phi(i \neq j)$。在同一子集$M_i$（或称等价类）中的各个元素，在一定意义上是不可区分的。如果把一个子集当成一个模式类，则满足等价关系的各类之间有明确的界限，或者说是可区分的。遗憾的是，相似关系不具有传递性。例如，父亲与儿子相似，儿子与母亲相似，但父亲与母亲未必相似。因此，实际的相似与分类问题远不像集合表达那样简单明了。集合的概念可以用来表现已经分好的类，但对于怎样分类和归类则缺乏指导意义。

目前得到广泛应用的相似性度量是在空间中定义的某种距离。给定一个输入样本集合X，我们用D维空间中的一个点表示某个样本，两个样本x_k和x_j之间的相似性度量$\delta(x_k, x_j)$应满足以下要求：

（1）相似性度量应为非负值，即$\delta(x_k, x_j) \geqslant 0$；

（2）样本本身之间的相似性度量应为最大；

（3）相似性度量应满足对称性，即$\delta(x_k, x_j) = \delta(x_j, x_k)$；

（4）在模式类满足紧致性的条件下，相似性应是点的间距的单调函数。

在各种空间中，只要定义一种距离度量，就可以用这种距离度量的非增函数作为相似性度量。例如，在D维欧几里德空间，可以选择某个随距离增加而下降的函数f作为相似性度量，即

$$\delta(x_k, x_j) = f\left[\sqrt{\sum_{i=1}^{D}(x_{ki} - x_{ji})^2}\right]$$

有些情况下，可以采用两个向量之间的夹角来度量相似性，例如

$$\delta(x_k, x_j) = \arccos \frac{x_k^{\mathrm{T}} x_j}{\|x_k\| \|x_j\|}$$

其中，x_k和x_j是表示模式k和模式j的向量，x_{ki}和x_{ji}是它们的第i个分量。

很显然，有了距离度量就可以判断相似性。如果样本集合满足紧致性条件且分类已知，就可以对样本进行归类。所以，模式识别研究的很大一部分工作是距离度量方法的研究。

在使用距离度量前，需要解决两个更深层的问题：

(1) 怎样决定分类；

(2) 怎样才能满足紧致性条件。

分类的复杂性，在于不存在纯客观的分类标准，因为任何分类都带有主观性。例如，鲸在生物学中属于哺乳动物，与牛可以分为一类，但从产业的角度，捕鲸属于水产业，和鱼是一类，而牛属于畜牧业。

靠哪些特征决定相似与否并进行分类呢？很大程度上依存于行为的目的和方法。因此在考虑分类特征和分类时，一定要明确分类的目的，并据此决定分类的方法。

3. 特征的选择与提取

特征是决定相似性与分类的关键。当分类的目的决定之后，如何找到合适的特征就成为识别的核心问题。从输入的原始信息得到特征，通常需要复杂的非线性运算，使用不同的方法，可以找到很多不同的特征。特征过多，会大大浪费模式识别的运算时间。特征选择与提取的基本任务是从许多特征中找出那些最有效的特征。

可以把特征分为三类：物理的、结构的和数学的。人们通常利用物理和结构特征来识别对象，因为这样的特征容易被视觉、触觉以及其他感觉器官所发现，但在使用计算机去构造识别系统时应用这些特征就比较复杂。计算机在抽取数学特征的能力方面比人强得多，因此，对计算机来说，更适合采用数学特征来进行识别。常用的数学特征有统计平均值、相关系数、协方差阵的本征值和本征向量等。

(1) 特征形成

根据被识别的对象产生出一组基本特征，它可以是计算出来的（当识别对象是波形或数字图像时），也可以是用仪表或传感器测量出来的（当识别对象是实物或某种过程时），这样产生出来的特征称为原始特征。

(2) 特征提取

原始特征的数量可能很大。为了便于处理，一般通过映射或变换，变成另外一组数量较少的特征，这个过程称为特征提取。映射后的特征称为二次特征，它们是原始特征的某种组合(通常是线性组合)。

(3) 特征选择

从一组特征中挑选出一些最有效的特征以进一步压缩特征的数量，这个过程称为特征选择。

例如，细胞自动识别。有一批包括正常及异常细胞的数字图像，要根据这些图像区分哪些细胞是正常的，哪些是异常的。首先找出一组能代表细胞性质的特征，如细胞总面积、总光密度、胞核面积、核浆比、细胞形状、核内纹理等，得到很多原始特征，然后从中选出对分类最有效的特征，即在正常细胞和异常细胞之间差别最大的特征。

再如，印刷体汉字识别。先扫描汉字的图像，再根据图像识别这是哪些汉字。首先找出一组和汉字结构特点有关的特征，如复杂指数、四边码、粗外围特征、粗网格特征、笔画密度特征、汉字特征点、包含配选特征、小笔段特征、差笔画特征等，然后从中选出对分类最有效的特征。

在选择了特征以后，就可以使用合适的分类方法进行分类决策，完成模式识别。

5.2.2　贝叶斯决策理论

模式识别的分类问题是指根据识别对象特征的观察值，将其分到某个类别中。统计决策理论是处理模式分类问题的基本理论之一。贝叶斯决策理论方法是统计模式识别中的一个基本方法，用这个方法进行分类时要求：

(1) 各类别总体的概率分布是已知的；

(2) 要决策分类的类别数是一定的。

5.2.2.1　基于最小错误率的贝叶斯决策

在模式分类问题中，人们往往希望尽量减少分类的错误。从这样的要求出发，利用概率论中的贝叶斯公式，就能得出使错误率最小的分类规则，称为基于最小错误率的贝叶斯决策。下面通过癌细胞的识别来说明这个方法。

假设每个要识别的细胞已做过预处理，抽取出d个表示细胞基本特性的特征，成为一个d维空间的向量$\boldsymbol{x}$，识别的目的是将$\boldsymbol{x}$分类为正常细胞或异常细胞。用ω表示细胞的状态。显然，细胞有两种状态ω_1和ω_2，用$\omega=\omega_1$表示正常，用$\omega=\omega_2$表示异常。

用某一地区正常细胞和异常细胞出现的比例作为先验概率，分别为$P(\omega_1)$和$P(\omega_2)$。则后验概率为

$$P(\omega_i \mid \boldsymbol{x})=\frac{p(x \mid \omega_i)P(\omega_i)}{\sum_{j=1}^{n} p(x \mid \omega_j)P(\omega_j)}$$

其中，n为状态数，这里只有两种状态，所以$n=2$。

基于最小错误率的贝叶斯决策规则为：如果$P(\omega_1 \mid \boldsymbol{x}) > P(\omega_2 \mid \boldsymbol{x})$，则把$\boldsymbol{x}$归类于正常状态；反之如果$P(\omega_1 \mid \boldsymbol{x}) < P(\omega_2 \mid \boldsymbol{x})$，则把$x$归类于异常状态。

例　假设在某个地方对人群细胞特征和患癌情况进行抽样调查，结果如表5.7所示。

表5.7　某地人群细胞特征和患癌情况

患者编号	细胞观察特征	跟踪结果	患者编号	细胞观察特征	跟踪结果
1	细胞核大小不一 细胞排列混乱	癌症	26	细胞核大小不一	
2	细胞排列混乱	癌症	27	细胞核大小不一	
3	细胞核大小不一		28	细胞排列混乱	
4	细胞核大小不一		29	细胞核大小不一	
5	细胞核大小不一 细胞排列混乱	癌症	30	细胞核大小不一	
6	细胞核大小不一		31	细胞核大小不一	
7		癌症	32		
8	细胞排列混乱	癌症	33	细胞核大小不一	
……			……		
25			50		

在表5.7中，细胞异常特征包括“细胞核大小不一”和“细胞排列混乱”，空白表示没有观察到异常特征。为简便起见，这里只使用了两个特征，实际上会有很多个特征。跟踪结果一列的空白表示该人正常，即没有患癌症。

细胞识别中，把细胞分为正常(ω_1)和异常(ω_2)两类，根据表5.7的数据，先验概率分别为$P(\omega_1)=0.9$，$P(\omega_2)=0.1$。现有一个识别的细胞，其观察值x=“细胞核大小不一”，根据类条件概率分布可得$p(x\mid\omega_1)=0.2$，$p(x\mid\omega_2)=0.4$，试对该细胞进行分类。

解：利用贝叶斯公式，分别计算出ω_1和ω_2的后验概率。

$$P(\omega_1\mid x)=\frac{p(x\mid\omega_1)P(\omega_1)}{\sum_j^2 p(x\mid\omega_j)P(\omega_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(\omega_2\mid x)=\frac{p(x\mid\omega_2)P(\omega_2)}{\sum_j^2 p(x\mid\omega_j)P(\omega_j)}=\frac{0.4\times0.1}{0.2\times0.9+0.4\times0.1}=0.182$$

根据贝叶斯决策规则，由于

$$P(\omega_1\mid x)=0.818>P(\omega_2\mid x)=0.182$$

所以，合理的决策是把x归类于正常状态。

当要分类的状态较多时，其计算与此类似，最终根据后验概率的计算结果，取后验概率最大的类别作为归类结果。表示如下：

如果$P(\omega_i\mid x)=\max\limits_{j=1,2,\ldots,n}P(\omega_j\mid x)$，则$x\in\omega_i$，其中$n$为状态数。

5.2.2.2 基于最小风险的贝叶斯决策

实际上，在决策时经常还需要考虑风险，风险和损失紧密相连，通常用如果决策错误则可能带来的损失大小来表示风险。

仍以癌细胞识别为例。对细胞的分类不仅要考虑到尽可能做出正确的判断，还要考虑做出错误判断时会带来什么后果。诊断中，如果把正常细胞判断为异常，则会给病人带来精神上的负担，而把异常细胞判断为正常则会延误病情，造成严重后果。显然，两种不同的错误判断所造成损失的严重程度不同，后者的损失比前者更严重。基于最小风险的贝叶斯决策正是考虑各种错误造成损失不同而提出的一种决策规则。

例 在前面癌细胞识别例子所给条件的基础上，设有如表5.8所示的决策表，按基于最小风险的贝叶斯决策进行分类。

表5.8 癌细胞识别的决策表

决策损失	状 态	
	正常(ω_1)	异常(ω_2)
正常(α_1)	0	6
异常(α_2)	1	0

解：已知条件为

$$P(\omega_1)=0.9,\qquad P(\omega_2)=0.1$$

$$p(x \mid \omega_1) = 0.2, \qquad p(x \mid \omega_2) = 0.4$$

$$\lambda_{11} = 0, \qquad \lambda_{12} = 6$$

$$\lambda_{21} = 1, \qquad \lambda_{22} = 0$$

从前例计算可知后验概率为

$$P(\omega_1 \mid x) = 0.818, \qquad P(\omega_2 \mid x) = 0.182$$

计算条件风险

$$R(\alpha_1 \mid x) = \sum_{j=1}^{2} \lambda_{1j} P(\omega_j \mid x) = \lambda_{12} P(\omega_2 \mid x) = 1.092$$

$$R(\alpha_2 \mid x) = \lambda_{21} P(\omega_1 \mid x) = 0.818$$

由于$R(\alpha_1 \mid x) > R(\alpha_2 \mid x)$，决策为$\omega_2$的条件风险小于决策为$\omega_1$的条件风险，所以判断待识别的细胞$x$为$\omega_2$类——异常细胞。

可以看出，由于决策规则不同，按照最小错误决策的结果和按照最小风险决策的结果可能并不相同。

5.2.2.3　其他决策规则

除了基于最小错误率的贝叶斯决策和基于最小风险的贝叶斯决策，还有一些其他的常用决策规则，主要包括：在限定一类错误率条件下使另一类错误率为最小的两类别决策、最小最大决策、序贯分类方法等。

1. 在限定一类错误率条件下使另一类错误率为最小的两类别决策

在两类别决策问题中，有犯两种错误分类的可能性。一种是在采取决策ω_1时其实际自然状态为ω_2，另一种是在采取决策ω_2时其实际自然状态为ω_1。实际中，有时要求限制其中某一类错误率不得大于某个常数而使另一类错误率尽可能地小。例如，在癌细胞识别中，我们已经认识到把异常判断为正常的损失更为严重，所以希望这种误判的错误率$P_2(e)$很小，在这种条件下，再要求$P_1(e)$尽可能地小。这种决策称为在限定一类错误率条件下使另一类错误率为最小的两类别决策，通常通过求条件极值的方法解决。

2. 最小最大决策

从最小错误率或最小风险贝叶斯决策中可以看出其决策都与先验概率$P(\omega_i)$有关。如果给定x，其$P(\omega_i)$不变，则按照贝叶斯决策规则，可以使错误率或风险最小。但如果$P(\omega_i)$是可变的，或者事先对先验概率毫无所知，若再按某个固定的$P(\omega_i)$条件下的决策规则来进行决策，则往往得不到最小错误率或最小风险。这时候可以选择使最小贝叶斯风险R^*为最大值时的$P^*(\omega_1)$来设计分类器，其最小贝叶斯风险的最大值$R^*_{\max}$相对于其他的$P(\omega_1)$为最大，从而能保证无论$P(\omega_1)$如何变化，使最大风险降为最小，这样的决策称为最小最大决策。

3. 序贯分类方法

前面所讲的方法中都认为d个特征都同时给出且不考虑获取特征所花的代价。在有些实际问题（如医疗诊断）中，特征的获取要花一定代价，这样除了考虑错分会造成分类损失以外，还应考虑获取特征所花的代价。可能会有这样的情况，获取了$k(k < d)$个特征后就能做到判决分类

更为合理。这是因为其余d-k个特征的加入使分类错误降低而造成的代价的减少补偿不了获取这些特征所花费的代价。

解决上述问题的方法可用序贯分类方法，即先用一部分特征来分类，逐步加入特征以减少分类损失。而每个步骤都要衡量加入新特征所花代价与所降低分类损失的大小，以便决定是否继续再加入新特征。

5.2.3 近邻法

近邻法是一类决策方法，它把各类中全部样本点用于设计分类方法（分类器）。常见的包括最近邻法、k-近邻法、可做拒绝决策的近邻法、最佳距离度量近邻法等。这里我们介绍其中两个比较简单的，即最近邻法和k-近邻法。

5.2.3.1 最近邻法

假定有c个类别$\omega_1, \omega_2, \cdots, \omega_c$的模式识别问题，每类有标明类别的样本$N_i$个，$i = 1, 2, \cdots, c$。

我们可以规定ω_i类的判别函数为

$$g_i(x) = \min_k \left\| x - x_i^k \right\|, \qquad k = 1, 2, \cdots, N_i$$

其中，x_i^k 的下标i表示ω_i类，上标k表示ω_i类N_i个样本中的第k个。决策规则可以写为

$$\text{若 } g_j(x) = \min_i g_i(x), \quad i = 1, 2, \cdots, c \qquad \text{则决策} x \in \omega_i$$

这一决策方法称为最近邻法。也就是说，对未知样本x，我们只要比较x与所有已知类别的样本之间的欧氏距离，与x距离最近的样本属于哪一类，就把x归入哪一类。

例 某事物可分成A、B、C三个类别，只选用一个特征来分类，每个类别有已知样本的特征值，如表5.9所示。

表5.9 已知类别的样本的特征值

类别	A	B	C
样本1	5	7	15
样本2	1	9	13
样本3		10	

现有一个待识别对象x，其特征值为5.5，用最近邻法对它进行分类。

解： 由于只选用了一个特征，定义其欧氏距离如下：

$$d = \sqrt{(x_1 - x_2)^2} = |x_1 - x_2|$$

计算x和所有已知分类样本的距离，结果如表5.10所示。

表5.10 x和已知类别的样本的距离

类别	A	B	C
样本1	0.5	2	10
样本2	4.5	4	8
样本3		5	

按照最近邻法的决策规则，和x距离最近的已知分类样本是A类中的样本1，因此把x归为A类。

5.2.3.2　k–近邻法

k-近邻法是最近邻法的推广，选取未知样本*x*的*k*个近邻，在这*k*个近邻中，属于哪一类的最多，就把*x*归为哪一类。

具体而言，是在*N*个已知样本中，找出*x*的*k*个近邻。设这*N*个样本中，来自ω_1类的样本有N_1个，来自ω_2的样本有N_2个，……，来自ω_c类的样本有N_c个，若$k_1, k_2, \cdots, k_c$分别是*k*个近邻中属于$\omega_1, \omega_2, \cdots, \omega_c$类的样本数，则可以定义判别函数为

$$g_i(x) = k_i, \qquad i = 1, 2, \cdots, c$$

决策规则为

$$\text{若 } g_j(x) = \max_i k_i \qquad \text{则决策} x \in \omega_j$$

例　某事物可分成A、B、C三个类别，只选用一个特征来分类，每个类别有已知样本的特征值，如表5.9所示。现有一个待识别对象*x*，其特征值为5.5，用*k*-近邻法对它进行分类。

解：由于只选用了一个特征，定义其欧氏距离如下：

$$d = \sqrt{(x_1 - x_2)^2} = |x_1 - x_2|$$

计算*x*和所有已知分类样本的距离，结果如表5.10所示。

按照*k*-近邻法，由于$k = 3$，所以从已知分类样本中选出3个和*x*距离最近的样本，分别为：A类的样本1，B类的样本1和样本2。在3个和*x*距离最近的已知分类样本中，属于B类的有2个，属于A类的有1个，属于B类的最多，所以把*x*归为B类。

5.2.4　印刷体汉字识别中的特征提取

汉字识别分为印刷体汉字识别和手写体汉字识别。这里以印刷体汉字识别为例，加深对模式识别中特征提取的理解。

在进行汉字识别前，先建立汉字特征库。首先把每个汉字的图形扫描到计算机中，成为图形文件，这种图形是由点阵组成的；然后针对得到的图形，提取其特征；最后根据提取的特征，对汉字进行分类，直到划分的每个类只包含一个元素。

进行汉字识别时，首先把页面扫描到计算机中，成为图形文件；然后把页面划分成一个个图形单元，每个图形单元是一个汉字的图形；针对每个图形单元，计算其特征，根据特征计算和每一个分类的距离，与哪个分类距离最小，就把它分为哪一类，也就是识别为哪一个汉字。

5.2.4.1　文字的归一化

把文字扫描到计算机中，在提取特征前，通常需要把文字进行归一化处理。归一化有位置归一化、大小归一化及笔画粗细归一化。这里只介绍位置归一化和大小归一化。

1. 位置归一化

为了消除汉字点阵位置上的偏差，需要把整个汉字点阵图形移动到规定的位置上，这个过程称为位置归一化。

有两种简单的位置归一化方法。一种是基于质心的位置归一化方法；另一种是基于文字外框的位置归一化。基于质心的位置归一化方法需要首先计算文字的质心，然后再把质心移动到指定的位置上。基于文字外框的位置归一化需要首先计算文字的外框，并找出中心，然后把文字中心移动到指定的位置上。

图5.4给出了两种方法的示例，从中不难看出，基于质心的位置归一化方法的抗干扰能力更强。

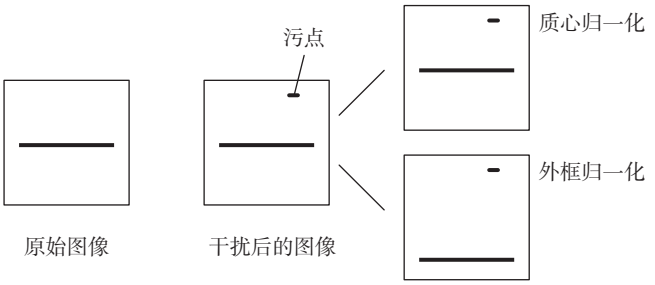


图5.4 噪声对位置归一化的影响

2. 大小归一化

对不同大小的文字进行变换，使之成为大小相同的文字，这个过程称为大小归一化。通过大小归一化，许多特征就能够用于识别不同字号混排的文字。

常用的大小归一化方法有两种。一种是将文字的外框按比例线性放大或缩小成为规定大小的文字。另一种是根据水平和垂直两个方向文字黑像素的分布进行大小归一化。

前一种方法只是个线性变换，不再赘述。对于后一种方法，首先计算文字的质心G_I和G_J：

$$G_I=\frac{\sum_{i=A}^{B}\sum_{j=L}^{R} i\cdot c(i,j)}{\sum_{i=A}^{B}\sum_{j=L}^{R} c(i,j)},\qquad G_J=\frac{\sum_{i=A}^{B}\sum_{j=L}^{R} j\cdot c(i,j)}{\sum_{i=A}^{B}\sum_{j=L}^{R} c(i,j)}$$

其中，$c(i,j)$的意义如下：$c(i,j)=1$表示该像素点为文字黑像素点；$c(i,j)=0$表示该像素点为背景。A、B、R和L分别表示文字的上下左右边界。

然后计算水平和垂直方向的散度σ_I和σ_J：

$$\sigma_I^2=\frac{\sum_{i=A}^{B}(\sum_{j=L}^{R} c(i,j))\cdot(i-G_I)^2}{\sum_{i=A}^{B}\sum_{j=L}^{R} c(i,j)},\qquad \sigma_J^2=\frac{\sum_{i=A}^{B}(\sum_{j=L}^{R} c(i,j))\cdot(j-G_I)^2}{\sum_{i=A}^{B}\sum_{j=L}^{R} c(i,j)}$$

一般计算出来的散度和规定的散度不同，最后按比例将文字线性放大或缩小成规定散度的点阵。

第一种方法容易受边框的噪声影响。第二种方法对于有些字，如“目”和“且”，归一化后会使得它们形状更相似而难以区别，如图5.5所示，其中(a)为基于外框的大小归一化结果，(b)为“且”字基于x、y方向离散度的大小归一化结果，(c)为“目”字x、y方向离散度的大小归一化结果。

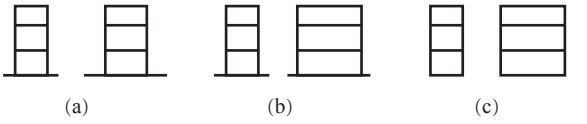


图5.5 两种大小归一化方法的比较

5.2.4.2 印刷体汉字识别中的一些特征

印刷体汉字识别中的关键问题是特征提取，需要研究哪些特征比较具有分类价值，同时又比较容易通过程序计算得到，然后使用这些特征来对汉字进行分类，即识别。用于识别的汉字特征应当对字体的不同、汉字大小的不同及噪声的影响等因素不敏感。

下面介绍一些印刷体汉字识别中的常用特征。

1. 复杂指数

文字x方向和y方向的复杂指数定义为

$$c_x = \frac{L_y}{\sigma_x}, \qquad c_y = \frac{L_x}{\sigma_y}$$

其中，c_x和c_y分别为x方向和y方向的复杂指数，L_x和L_y分别为x方向和y方向黑像素的总数，σ_x和σ_y分别为x方向和y方向质心二次矩的平方根：

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(i-G_I)^2 c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)}}, \qquad \sigma_y = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(j-G_J)^2 c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)}}$$

其中，G_I和G_J分别为文字质心位置的i和j坐标值，N和M是文字点阵的长和宽。

复杂指数反映了文字在x方向和y方向笔画的复杂程度。该特征对文字的位置和大小不敏感。

2. 四边码

从文字四周边框开始，向内取适当宽度，以此宽度分割出文字四周的四个部分。根据每个部分中含有文字黑像素的多少分为四级编码（0, 1, 2, 3）。如图5.6中所示的“昨”，其四边码为“0102”。

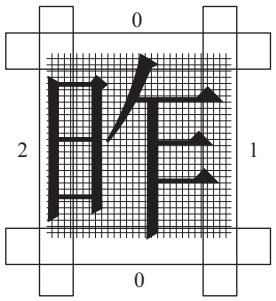


图5.6　文字四边码举例

四边码特征对文字的断线有较强的适应性。

3. 粗外围特征

粗外围特征抽取的过程为：把$p \times q$点阵文字分割成$n \times n$份，n通常取8。从文字四边框各向对边扫描，计算最初与文字笔画相碰的非文字部分的面积和全部面积之比，将其作为一次粗外围特征p_{1i}(i = 1, …, 4n)，再将第二次与文字点相碰的非文字部分面积和全部文字面积之比作为二次粗外围特征p_{2i}(i = 1, …, 4n)，形成8n维的特征向量$\boldsymbol{p}$（见图5.7）。

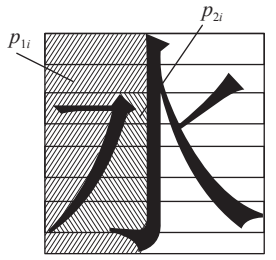


图5.7　粗外围特征

一次粗外围特征反映了文字轮廓特征，二次粗外围特征在某种程度上反映了文字内部结构。

4. 粗网格特征

把$p \times q$点阵文字分割成$n \times n$份，n通常取8，取每份文字中的黑像素数与整个文字黑像素数之比，将所有$n \times n$值排成一列，形成n^2维特征向量。

粗网格特征体现了文字整体形状的分布，但该特征抗笔画位置干扰的能力差。

5. 笔画密度特征

在$p \times q$点阵中，向不同的方向投影，对文字黑像素的个数进行累加计算，形成笔画密度直方图。通常取水平、垂直、45° 和135° 四个扫描方向，每个方向取n（通常n = 16）个值作为特征，形成4n维特征向量（见图5.8）。

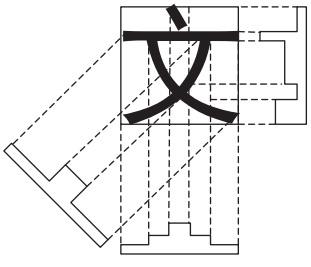


图5.8　笔画密度特征

这种从文字四个方向抽取的笔画密度特征称为四方向笔画密度特征。它不但对印刷体汉字分类有较好的效果，对手写体汉字分类也具有价值。

6. 汉字特征点

一个汉字的笔画和背景中的关键点是汉字结构的一种重要字形特征。

汉字基本由直线笔画构成，是一种直线型文字。在一幅二值化图像中，汉字信息绝大部分集中在汉字骨架上，而汉字骨架信息又大多集中在若干特征点，称为笔画特征点，如图5.9所示。一旦确定笔画特征点，根据若干连接规则，就可以确定汉字笔画以及结构形状。

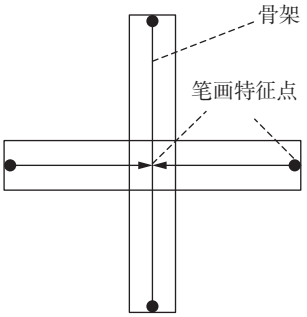


图5.9 汉字笔画的骨架和特征点

一个汉字图形的背景部分也包含了区别于其他汉字的丰富信息。在背景部分选取关键点（称为关键背景点），也可以有效地区分汉字。对笔画少的汉字，选取关键背景点尤其重要。

汉字笔画特征点可以选取端点、折点、歧点和交点。端点是笔画的起点或终点且不与别的笔画相接；折点是笔画方向出现显著变化的点；歧点是三叉点，要求其中两个笔段的分支方向相同；交点是四叉点且有两对相等的对顶角，如图5.10所示。

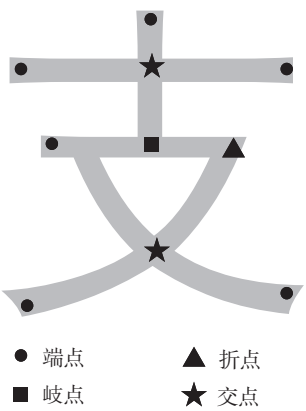


图5.10 汉字的笔画特征点示例

汉字笔画特征点集中了主要的汉字结构信息，无论是北方印刷汉字还是南方印刷汉字，无论是书版还是报版字，无论是宋体印刷汉字还是其他体印刷汉字，甚至是书写规整的手写体汉字，同一汉字的特征点非常稳定，其中折点、交点尤其稳定。因此，使用汉字特征点，原理上就可以识别多体印刷汉字，甚至可以识别手写体汉字。

汉字特征点反映了汉字结构特征。与统计特征相比，汉字中的非结构信息（如笔画粗细、字形位置变动、少量旋转等）的不稳定性，从理论上讲，对汉字特征点的提取无影响。所以，用特征点来识别汉字，可以增加抗噪声能力，提高实用性。

通常情况下，要提取笔画的特征点，首先要对文字图像进行处理，将文字的笔画变细，这个过程称为细化。该方法对图像处理技术提出了较高要求，如果文字图像处理得不够理想，会影响到该方法的效果。

7. 包含配选法

许多汉字具有相同的偏旁部首，包含配选法就是利用这一特点对汉字分类的。分类用的模板是汉字偏旁部首的骨架图形。分类时，将输入文字和各标准模板做“与”运算。显然，只有和输入未知文字的偏旁部首相同的标准模板相“与”，其结果才和标准模板本身的图形一致，如图5.11所示。所以，根据未知输入文字图像和分类用标准模板图像相“与”的结果是否和该标准图像相同，可以判别出未知文字属于哪一类。

(a)标准模板 (b)输入文字 (c)相“与”结果

图5.11 包含配选法原理

在没有和标准图像相“与”匹配前，先把未知文字图像横线加粗成大于等于3个像素宽，以利于包含相匹配的模板，但是这样也加大了包含其他标准模板的概率，因而误识率会增加。为避免文字笔画绝对位置移动带

来的干扰，可以把图像沿上下左右四个方向平移一个像素，然后分别与标准模板相匹配。只要有一次匹配成功，就判定该文字属于标准模板图像的类别。包含配选法实际上就是模板匹配法。

当识别字数增多时，由于偏旁部首的标准模板增加很少，其分类速度比粗外围和粗网格法更容易提高，而存储量要求较少。

8. 基于小笔段特征的层次结构

汉字的笔画特征受字体、字号等影响较小，是识别汉字的很好的特征。但是汉字笔画特征对实际文本来说很难提取，用基于小笔段特征的汉字层次结构，能较好地解决以上问题。

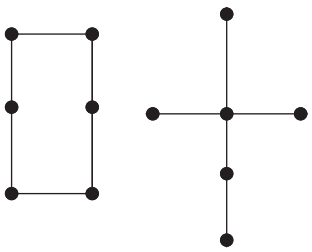


图5.12　汉字“叶”的小笔段表示

若干小笔段首尾相连构成了汉字笔画，如图5.12所示。用小笔段作为基元，一方面易于提取，另一方面又保留了汉字基本的笔画结构信息，并且在字体变化或噪声干扰条件下，仍能保持笔画结构的绝大部分信息。

小笔段组成了部件，部件又组成了汉字，由小笔段到部件，再到汉字的层次结构描述，反映了汉字结构不同层次的约束关系。

汉字的字体改变和干扰影响会使得小笔段特征向量发生变化。因此，用层次结构法对未知汉字进行匹配判别时，采用精确匹配方法往往不能奏效。较好的办法是采用由汉字小笔段相关系数约束的松弛匹配算法。这种算法能有效地吸收同一汉字不同字体的变化，而所能容许的变化范围由小笔段间的相关系数所制约。实验证明，这种特征和算法可以有效地解决多体印刷汉字识别的问题。

9. 差笔画特征

粗分类后，通常每类中含有许多形状相似的文字，采用差笔画方法，可以较好地区分这些形状相似的汉字。差笔画法是一种适用于多体印刷汉字细分的方法。差笔画细分的算法如下：

（1）设粗分类后，类中只有两个文字A和B。

（2）预先根据文字A和B特征点（如端点、折点等）坐标生成其骨架图形S_A和S_B，如图5.13所示。

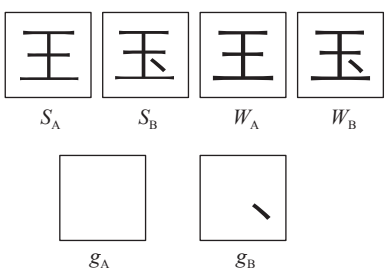


图5.13　差笔画模板示例

（3）把S_A和S_B在3×3网格区域内移动位置，生成如图5.13所示的笔画宽度为3个像素的粗图形W_A和W_B。

（4）对S_A第i号笔画S_{A_i}在5×5网格区域内挪动，与W_B匹配，通过下式求出最大一致度σ_{max}：

$$\sigma_{\max}=\max_{5\times5}\frac{(S_{A_i}\cdot W_B)}{\|S_{A_i}\|}$$

求出$s_{max}\leqslant\theta_0$（θ_0为常数）的笔画，作为差笔画g_A。

（5）同样由S_B和W_A求出差笔画g_B。g_A和g_B可能同时存在，也可能仅有一个。

（6）若差笔画仅有一个（如g_A）时，把输入文字图形在3×3网格区域内与存在的差笔画g_A进行位置匹配，由下式求出最大一致度σ_A：

$$\sigma_A=\max_{3\times3}\frac{(x\cdot g_A)}{\|g_A\|}$$

若$\sigma_A\geqslant\theta_0$（$\theta_0$为常数），则该图形属于A，否则属于B。

(7) 若两个差笔画g_A和g_B都存在，在3×3网格区域内将输入图形与g_A和g_B进行位置匹配，求出各自的最大一致度σ_A和σ_B。若$\sigma_A \geqslant \theta_1$且$\sigma_A - \sigma_B \geqslant \theta_2$，则$x \in A$；若$\sigma_B \geqslant \theta_1$且$\sigma_B - \sigma_A \geqslant \theta_2$，则$x \in B$。其中，$\theta_1$和$\theta_2$为常数。

(8) 若粗分类后，类中文字大于两个时，则先取两个来进行上述算法的判别，以后反复取两个进行判别，来决定输入文字属于的类别。

5.2.4.3 分类

由于汉字数量大，所以分类的计算量很大。为了减少计算量，提高识别效率，一般对汉字进行一级或多级分类，即先粗分类，然后再细分类，直到识别出汉字。对于粗分类，要求正确分类率和分类稳定性要高，分类的速度要快，用来进行粗分类的特征要简单。

下面介绍两种分类方法。

1. 采用复合特征的分类

选用N种具有互补特性的特征作为类特征。在学习阶段，可对训练样本进行N次互不相关的分类，然后组合N次分类结果，完成特征空间的划分。分类时，根据待分字的特征进行N次分类，组合分类结果，以求得子类。组合的形式有两种：

$$S = S_1 \cup S_2 \cup \cdots \cup S_N$$

$$S = S_1 \cap S_2 \cap \cdots \cap S_N$$

其中S_N表示第N次分类的子类。有人采用第一种组合形式，其第一种特征是网格特征，第二种特征是周边特征。有人采用第二种组合形式，其第一种特征是复杂指数，第二种特征是四边码。

2. 多级分类

在学习阶段，可对训练样本进行多级分类，每一级分类是在上级分类基础上进行的，分类时重复上述多级分类过程。有人采用决策分类树分类的方法，用Walsh变换系数为特征，用模糊搜索为策略进行树分类。也有人采用4×4、8×8、16×16、32×32这四种标准模板，把输入的待识别文字首先进行4×4图形匹配，把差别小于一定阈值的字保留，再顺序用8×8、16×16、32×32的模板以同样的方式进行分类，最后只取相似的一个字作为识别结果。

5.2.4.4 距离度量

在选取了特征之后，需要选择或寻找适当的判别准则，从而判断出待识别文字的特征与哪一个类别的特征最近。常用的距离度量准则如下：

(1) $D(\boldsymbol{X}, \boldsymbol{G}) = \sum_{i-1}^{m} |x_i - g_i|$

(2) $D(\boldsymbol{X}, \boldsymbol{G}) = \left[(\boldsymbol{X} - \boldsymbol{G})^{\mathrm{T}} \sum^{-1} (\boldsymbol{X} - \boldsymbol{G}) \right]^{1/2}$

(3) 相似度

$$R(\boldsymbol{X}, \boldsymbol{G}) = \frac{(\boldsymbol{X} \cdot \boldsymbol{G})}{\|\boldsymbol{X}\| \cdot \|\boldsymbol{G}\|} = \frac{\sum_{i=1}^{m} x_i \cdot g_i}{\left[\sum_{i=1}^{m} x_i^2 \cdot \sum_{i=1}^{m} g_i^2 \right]^{1/2}}$$

以上各式中，$\boldsymbol{X}$、$\boldsymbol{G}$都是特征向量，分别表示待识别文字的特征向量和标准类别特征向量，x_i和g_i分别是$\boldsymbol{X}$和$\boldsymbol{G}$的分量。

5.3　信息融合技术

信息融合是指将来自多传感器或多源的数据进行协调优化和综合处理，产生新的有价值的信息，以得出更为准确、可信的结论。信息融合可以分成三个级别，数据级信息融合、特征级信息融合和决策级信息融合。本节将介绍这三级信息融合的基本概念与原理。

5.3.1　数据级信息融合

数据级信息融是指将来自多个传感器的原始数据直接进行融合。其原理如图5.14所示。各传感器获取数据以后，首先对原始数据进行关联和配准，然后进行信息融合，得到质量更高的信息。

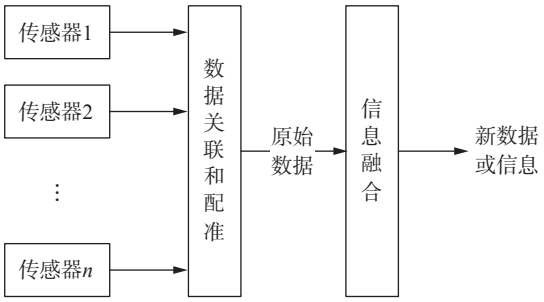


图5.14　数据级融合示意图

数据级融合主要应用于图像融合，所以有些文献也把数据级融合称为像素级融合。图像融合是将不同传感器获得的同一景物的图像配准后，合成一幅新图像，以克服各单一传感器图像在几何、光谱和空间分辨率等方面存在的局限性。数据级融合还应用于同类型（同质）雷达波的直接合成，以改善雷达信号处理的性能。例如，合成孔径雷达就使用了数据级信息融合来增强信号，提高发现目标的能力。

照相是重要的军事侦察手段，随着传感器技术的发展，能够获得的图像已经不限于可见光图像，还可以得到热红外图像、微波图像等。可见光图像分辨率较高，符合人们的感觉习惯；热红外图像分辨率较低，得到的图像与人们的认知有所不同，有时候有些模糊难辨，但是可以识破多种伪装；微波图像分辨率较低，但是可以穿透一定的障碍物。这些图像各有所长，如果能够对同一地区获取可见光、热红外和微波三种图像，把它们融合在一起，就可以得到更加丰富的信息，比如可以识别和定位隐藏的目标。

5.3.1.1　图像融合示例

图5.15所示的是对同一目标区域用不同的设备拍摄的两幅图像。左边的图像是用微光夜视设备拍摄的微光夜视图像，相对比较清晰，可以看出这是一片茂密的树林。右边的图像是用热像仪拍摄的热红外图像，由于人体温度比周围植物的温度高得多，所以在热红外图像中，人体显得很明亮。但周围的植物温度相近，对比模糊，难以分辨，所以对人周围的环境情况看不清楚。如果把这两幅图像融合在一起，就可以清楚地看出人隐藏在什么位置。图5.16是融合后的图像，可以清楚地看到一个人隐藏在树林中。

图5.15 同一目标区域的微光夜视图像和热红外图像

图5.16 图像融合后的效果

5.3.1.2 图像融合原理

图像融合主要包括空间配准和图像融合两个步骤，如图5.17所示。

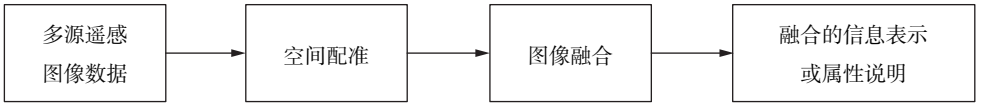


图5.17 多源遥感图像融合的一般模型

1. 空间配准

要把两幅图像融合在一起，首先必须找到两幅图像的像素的对应关系。两幅来自不同传感器的图像，大小、角度等一般都会有所不同，其像素并不是一一对应的。为使两幅图像能够对应起来，需要进行专门的处理，这个处理过程称为空间配准。

图像的空间配准一般是把其中一幅作为参考图像，即以它为基准，对另一幅图像进行校正，可分为如下4个步骤。

(1) 在两幅图像上，选择明显的特征，如边界、线状物交叉点、区域轮廓线等；

(2) 找出两幅图像上对应的明显物点，作为控制点；

(3) 根据控制点，建立图像之间的映射关系；

(4) 根据映射关系，对非参考图像进行重采样，获得与参考图像配准的图像。

2. 图像融合

将图像进行空间配准以后，就可以用合适的融合算法把两幅图像融合在一起。从两幅原图可以看出，可见光图像比较清晰，图像层次丰富，而热红外图像则整体亮度很低，只有人体呈现出很高的亮度。显然，如果把热红外图像的亮度增加到可见光图像中，就可以在保留可见光图像细节的前提下显示出人的位置。

在计算机中，颜色的表达一般使用RGB模型，即用红、绿、蓝三原色的各种组合来表示各种不同的颜色。红、绿、蓝三个分量的取值范围都是0~255，可以表达1600多万种颜色。RGB色彩模型非常适合硬件处理。

另一种常用的色彩模型是HSI模型，它使用色调、饱和度和亮度三个分量来表达颜色，比RGB色彩模型更符合人眼的感觉规律。

RGB模型和HSI模型可以相互转换。首先把R、G和B的值归一化到 [0, 1] 区间内，然后可以用如下公式将用RGB模型表示的数据转换成HSI模型表示的形式：

$$H=\begin{cases}\theta, & B\leqslant\theta\\ 360-\theta, & B>G\end{cases}$$

其中，

$$\theta=\arccos\left(\frac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right)$$

$$S=1-\frac{3}{(R+G+B)}(\min(R,G,B))$$

$$I=\frac{1}{3}(R+G+B)$$

将HSI模型表示的数据转换成RGB模型表示，需要分三种情况。

当0° ≤H<120° 时：

$$B=I(1-S)$$

$$R=I\left(1+\frac{S\cos H}{\cos(60°-H)}\right)$$

$$G=3I-(R+B)$$

当120° ≤H<240° 时，首先从H中减去120°，即$H=H-120°$，则有

$$R=I(1-S)$$

$$G=I\left(1+\frac{S\cos H}{\cos(60°-H)}\right)$$

$$B=3I-(R+G)$$

当240° ≤H<360° 时，首先从H中减去240°，即$H=H-240°$，则有

$$G=I(1-S)$$

$$B=I\left(1+\frac{S\cos H}{\cos(60°-H)}\right)$$

$$R=3I-(G+B)$$

这时得到的RGB值在 [0, 1] 区间内，再乘以值域，就得到了所需的RGB值。

要把热红外图像的亮度融合到可见光图像中，可以把两幅图像转换成HSI格式，然后把热红外图像的亮度增加到可见光图像对应点的I，例如：

$$I_{融}=0.6I_{可}+0.4I_{热}$$

其中，$I_{融}$为融合后的图像点亮度，$I_{可}$为对应的可见光图像点亮度，$I_{热}$为对应的热红外图像点亮度。

考虑到热红外图像中主要是亮度高的部分为有效信息，为了避免亮度低的部分影响图像清晰度，可以增加一个亮度阈值，只把超过亮度阈值的热红外图像像素融合到可见光图像中，即

$$I_{融}=\begin{cases}I_{可}, & I_{热}<I_{阈}\\ 0.6I_{可}+0.4I_{热}, & I_{热}\geqslant I_{阈}\end{cases}$$

其中，$I_{阈}$为执行信息融合的热红外图像像素亮度阈值。然后再把HSI格式的图像转换成RGB格式，就完成了信息融合，即由图5.15所示的两幅图像得到图5.16所示的融合图像。

图像的融合既可以通过软件实现，也可以通过硬件实现。前面的示例是一种非常简单的计算方法，在实际中，图像融合的计算方法要复杂得多。常用的融合方法有：代数法、HSI变换法、高通滤波法、回归模型法、PCA变换法、卡尔曼滤波法、小波变换法等。在实际工作中，需要根据图像的特点和图像融合的目的，采用合适的方法。

5.3.1.3 图像融合效果评价

数据级图像融合效果的评价分为主观评价和客观评价，一般结合起来使用。主观评价是指通过目视效果进行分析；客观评价是指利用图像的统计参数进行判定。

融合图像的客观评价应符合主观评价，也就是说，图像的统计参数特征应该符合人眼的目视感觉。遥感图像信息融合具有特殊性，它不仅要求提高融合图像的空间分辨率，而且要求尽可能地保持原始图像的光谱特征，而这两个要求在很大程度上是不相容的。因此，对于遥感图像融合效果的评价，应综合考虑空间细节信息的增强与光谱信息的保持。一般应综合利用三类统计参数来进行分析与评价：第一类反映亮度信息，如均值；第二类反映空间细节信息，如方差、信息熵和清晰度；第三类反映光谱信息，如扭曲程度、偏差指数与相关系数。下面简要介绍各种参数的定义及其物理含义。

1. 均值与标准方差

在统计理论中，统计均值μ和标准方差σ^2定义为

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i\ ,\qquad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

其中，n为样本总数，x_i为第i个样本值。

对某一幅图像，n为像素总数，x_i为第i个像素的灰度值，则均值为像素的灰度平均值，对人眼反映为平均亮度。如果均值适中（灰度值在128左右），则视觉效果良好。方差反映了灰度相对于灰度均值的离散情况，方差越大，则灰度级分布越分散。

2. 信息熵

对于一幅单独的图像，若其各元素的灰度值是相互独立的样本，则这幅图像的灰度分布为$P=\{p_0, p_1, \cdots, p_{L-1}\}$，$p_i$为灰度值等于$i$的像素数与图像总像素数之比，$L$为灰度级总数。对于灰度范围为$\{0, 1, \cdots, L-1\}$的图像直方图，其信息熵定义为

$$H=-\sum_{i=1}^{n}p_i\ln p_i$$

显然，$0\leqslant H\leqslant\ln L$。当某个$p_i=1$时，$H=0$；当$p_0=p_1=\cdots=p_{L-1}=1/L$时，$H=\ln L$。

图像信息熵是衡量图像信息丰富程度的一个重要指标。熵的大小反映了图像携带的信息量的多少。融合图像的熵值越大，说明其携带的信息量越大，图像的细节表现能力越强。图像中所有灰度级的出现概率越趋于相等，则包含的信息量越趋于最大。

3. 清晰度

图像清晰度采用梯度法来衡量，图像的梯度计算公式为

$$g=\frac{1}{n}\sum\sqrt{(\Delta I_x^2+\Delta I_y^2)/2}$$

其中，ΔI_x和ΔI_y分别为x与y方向上的差分；n为图像的大小。如果g越大，则图像的清晰度越高。很多情况下，图像融合可以有效地改善清晰度。

4. 扭曲程度

图像光谱扭曲程度直接反映了多光谱图像的光谱失真程度。光谱扭曲定义为

$$D=\frac{1}{n}\sum_i\sum_j\left|I'(i,j)-I(i,j)\right|$$

其中，n为图像的大小；$I'(i,j)$与$I(i,j)$分别为融合后和原始图像上 (i,j) 点的强度值。

5. 偏差指数

偏差指数由融合前的强度I和融合后的强度值I'经下式计算得到，即

$$D_{\text{index}}=\frac{1}{n}\sum_i\sum_j\frac{\left|I(i,j)-I'(i,j)\right|}{I(i,j)}$$

偏差指数用来比较融合图像和低分辨率多光谱图像的偏离程度。

5.3.2　特征级信息融合

特征级信息融合是指先对各个传感器的原始信息进行特征提取，例如提取出目标的边缘、方向、距离、速度等特征信息，然后对这些特征信息进行综合分析和融合处理，得到融合后的特征。其原理如图5.18所示。各传感器获取目标的特征信息后，在时间和空间上进行配准，然后进行数据关联和状态估计，得到目标的航迹和状态。目标既可能是一个，也可能是多个，目标航迹和状态也可能是多个。

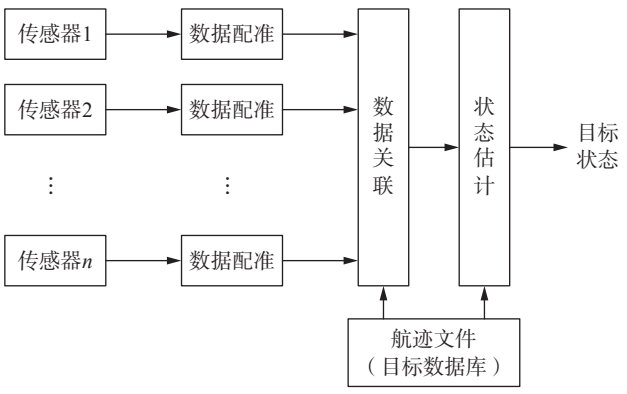


图5.18　特征级目标状态融合

特征级融合从大量原始信息中提取出少量特征信息，实际上实现了信息压缩，所以有利于实时处理。特征级信息融合的最典型实例就是组网雷达的信息融合。

雷达是重要的信息获取装备。早期的雷达主要用来探测飞行器，并测量飞行器的距离、方位、速度等特征数据。随着技术的发展，也出现了可以成像的雷达，如合成孔径雷达和逆合成孔径雷达。然而最常用的雷达，仍然主要用来探测目标的几个特征数据。

随着反辐射导弹技术、装备隐身技术和电子干扰技术的发展，雷达技术面临着严峻的考验。反辐射导弹可以追踪雷达天线射出的电磁波束，沿着电磁波束攻击雷达天线，先进的反辐射导弹还采用了人工智能技术，使反辐射导弹具有记忆功能，能对关机之后的雷达进行攻击，并能自动切换制导方式，自动搜索和截获目标，这大大提高了目标攻击的准确性和杀伤力，对雷达形成极大的威胁。装备隐身技术可以使装备的迎面雷达截面积显著降低，使得传统预警雷达探测目标的距离大大缩短，无法为防卫武器提供足够的预警时间，而火控雷达可能无法锁定目标，使武器系统失去自动攻击能力。电子干扰会严重扰乱雷达的信号，使其无法准确探测目

标。现代干扰设备中采用了先进的综合电子干扰系统、多功能多波束相控阵干扰系统和智能技术，给雷达造成了更严重、更致命的威胁。

为了应对这些威胁，雷达组网技术得到了发展。组网的雷达，不仅生存能力得到了很大提高，其探测能力、探测精度也大大提高了。组网雷达系统要求来自多个雷达的信息得到有效的融合处理，其融合对象是多个雷达探测到的目标的距离、方位、速度等特征数据，因此这种信息融合也称为基于特征的信息融合。

参与组网的不仅有雷达，也包括电子支援测量（ESM）系统。所谓ESM系统，实际上就是一个精密的电磁辐射接收与分析系统，它自身并不发射电磁波，但可以像雷达那样用天线接收一定方向上的电磁辐射源发射的电磁波，分析并记录其频谱和波形等特征，通过分析判断辐射源的特性，实现对目标的识别和威胁等级判断。飞机等装备上都有雷达、电台等电磁辐射源，ESM系统可以探测、分析和记录其特征，作为识别装备身份的依据。电子支援测量系统和雷达一起，构成了一个战场多元信息融合处理系统。

ESM系统的探测范围比雷达大，这在一定程度上扩大了组网系统的监视范围，但它一般不具备测距能力；雷达是一种有源传感器，通过发射电磁波，利用目标的回波进行目标探测，具有较强的定位能力，但对目标辐射源的技术性能和工作状态无法进行准确的区分，对目标的识别能力较弱。通过雷达获取目标的方位和距离信息，通过ESM系统获取目标的方位和辐射源识别信息，两者结合共同完成对目标的定位、识别和威胁等级判断，是组网雷达系统的一种发展趋势。

通过雷达组网系统进行目标识别和威胁等级判断，通常需要三个步骤：雷达与ESM航迹关联、目标融合识别、目标威胁判断。

1. 雷达与ESM航迹关联

雷达与ESM数据的关联分析是两者信息融合的前提和关键。假设由若干部雷达和ESM设备共同对某一监视区域内的多个目标进行跟踪监视，雷达和ESM的覆盖区域相互重叠。若雷达已探测到m条跟踪航迹，则分别与m个平台相关联。ESM也按照时序$t_i(i = 1, 2, \cdots, n$，其中n为测向次数)对辐射源目标进行测向，获得一组测向数据$b(t_i)$，构成一条ESM航迹。数据关联就是要确定这组测向数据是否与已知雷达航迹关联，与哪一条航迹关联。雷达的一条航迹对应一个目标，一条ESM航迹对应一个辐射源，一个目标可以搭载多个辐射源，因此，雷达的一条目标航迹可以和多条ESM航迹关联，一条ESM航迹至多和一条雷达航迹关联。

关联分析一般通过关联判别模型进行。在进行检验时，需要首先对数据进行时间校准，由于雷达的观测精度远比ESM测向精度高，所以通常以雷达观测为准将雷达航迹外推到ESM观测时间；然后，通过判别模型分析雷达航迹和ESM航迹的关联度或相似度；再后，按照一定的关联判决规则，把雷达航迹和ESM航迹关联起来。基于不同的情况，常用的关联判决规则有硬判决规则、双门限判决规则、三门限判决规则和四门限判决规则等。

2. 目标融合识别

现代战争的电磁环境日益复杂，组网雷达的目标识别能力受到多种因素的影响，使目标识别具有很大的不确定性。为了提高组网雷达系统的目标识别能力，可以将多个传感器（包括雷达和ESM等传感器）提供的目标识别信息通过证据理论进行有效融合，消除不确定性，获得可信度较高的目标识别结果。

由于雷达和ESM具有不同的观测周期且用于目标识别的特征不同，因此一般采用决策级融合。融合分为时域融合和空域融合，雷达和ESM传感器在各自的观测周期上先进行时域融合，得到各自的目标识别结果，然后再在融合中心进行空域上的决策级融合。

通常分别在时域和空域上定义基本概率分配函数，该分配函数用来确定雷达和ESM的观测结果分别在时域和空域上对各判断结果的支持力度，判定的目标类型应具有最大的基本概率分配函数值，且与其他目标类型的基本概率分配函数值之差大于某个门限，且满足判定的目标类型基本概率分配函数值大于不确定性基本概率分配函数值。

3. 目标威胁判断

目标威胁判断是组网雷达系统的一项重要工作内容，是对敌方目标对我方重要目标存在潜在威胁的判断。根据武器装备的反应时间和雷达特定的工作环境，可以将目标的威胁程度分为紧急目标、严重威胁目标、严重目标、一般目标和安全目标五种类型。

目标威胁判断是指根据雷达和ESM探测到的信息，把多个目标按照一定的规则分别划分为紧急目标、严重威胁目标、严重目标、一般目标和安全目标五种类型，如图5.19所示，其中d_i为目标到我军重要目标的距离，t_i为目标到达我军重要目标的时间。

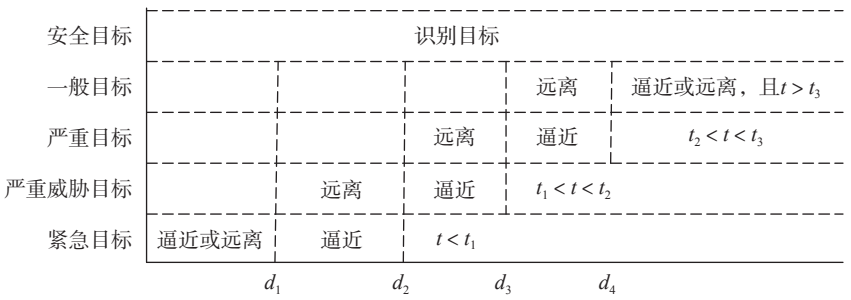


图5.19　威胁判断示意图

例如，在图5.19中，属于紧急目标的包括：(1) 融合中心不能识别，目标距我方重要目标的距离小于d_1；(2) 融合中心不能识别，目标距我方重要目标的距离区间为（d_1, d_2），且目标逼近我方重要目标；(3) 融合中心不能识别，目标距我方重要目标的距离大于d_2，且目标到达我方重要目标的时间小于t_1。

5.3.3　决策级信息融合

决策级信息融合就是多分类器融合，也称多分类器集成，是指融合多个分类器提供的信息，得到更加精确的分类（识别）结果。这是一种高层次融合，其原理框图如图5.20所示。每个传感器完成本地处理，包括预处理、特征提取、识别或判决，得到所观察目标的初步决策；然后，通过关联处理，保证参与融合的决策是源于同一观察目标的；最后，进行决策的融合判决，获得联合推断结果。决策级融合的主要方法包括贝叶斯推理、D-S证据理论、模糊集合论、神经网络、专家系统方法等。

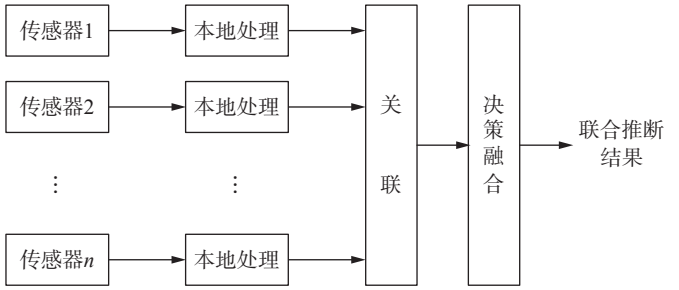


图5.20　决策级融合

决策级融合具有灵活性高、通信量小、容错性强、抗干扰能力强、对传感器的依赖性小（同质或异质）、融合中心处理代价低等优点。但是，由于需要先对各个传感器数据进行预处理以获得各自的判决结果，所以预处理代价较高。

之所以需要多分类器融合，主要有以下两方面的原因。

（1）不同分类方法的互补性。不同的分类器采用不同的分类方法，其各有优点，但有一种方法能适合所有的应用需要，或达到期望的效果。不同的分类器对于分类模式具有互补的信息，可以利用这些信息提高识别和分类性能。

（2）复杂模式识别的需要。对于复杂的模式识别问题，输入的特征变量多，类型和表现形式各异，量级不同，难以用一个分类器来处理。利用多分类器融合技术，可以降低对每一个分类器的性能要求，降低总体设计难度。

5.3.3.1 多分类器信息融合分类

给定模式空间S，由M个互不相交的模式类集合$\omega_1, \omega_2, \cdots, \omega_M$组成，即$S = \omega_1 \cup \omega_2 \cup \cdots \cup \omega_M$，$\omega_i \cup \omega_j = \Phi(i \neq j;\ i, j = 1, 2, \cdots, M)$，模式识别的目标是把给定的模式$x$划分到$\omega_1, \omega_2, \cdots, \omega_M$中的一个。

设e为一个分类器，令$A = \{1, 2, \cdots, M\}$。对于输入样本（模式）$x \in S$，$e(x) = j$表示分类器e把x划分到类ω_j中。其中，$j \in A \cup \{M + 1\}$；j为模式类集合ω_j的类别号（标签）；$e(x) = M + 1$表示分类器e拒识x。

许多分类器不仅能够提供类别号j，还能够提供其他的相关信息。一般地，分类器能够提供的信息可分为三个级别：抽象级、排序级和度量级。

在这三个级别中，度量级包含的信息最多，排序级次之，抽象级信息最少。从度量级到抽象级是一个信息减少的过程或抽象的过程。根据度量级信息可以得到排序级、抽象级信息，根据排序级信息可以得到抽象级信息。例如，根据属于某个标签的度量值，按照某种排序规则，可以把A中的标签排成一个序列（递增或递减）；通过选择顶层的标签，或直接在度量级选择具有最大值或最小值的标签，分配给输入样本x。

许多分类器能够提供度量级信息，如贝叶斯分类器中的后验概率、距离分类器中的距离、模糊分类器中的隶属度等。有些分类器只能提供抽象级输出，如句法分类器。

对应于分类器信息的级别，多分类器融合可以分为3种类型，即决策级融合、排序级融合和度量级融合。这三类方法覆盖了不同的应用范围，所利用的分类器输出信息量依次增多，也可能相应地得到更好的结果。

1. 抽象级融合

在抽象级融合中，分类器e只输出唯一的标签j，即某个确定的类别号，或一个子集$J \subseteq A$。

假设R个分类器$e_1, e_2, \cdots, e_R$对同一输入样本（模式）x进行分类，事件$e_k(x) = j_k$表示分类器e_k把x划分到类ω_j中，其中，$j_k \in A \cup \{M + 1\}$，$e_k(x) = M + 1$表示分类器e_k拒识x，$k = 1, 2, \cdots, R$。决策级融合是指利用这些事件，构造一个集成的分类器E对x进行分类，输出一个确定的类别号，即$E(x) = j$，$j \in A \cup \{M + 1\}$。

任何排序级和度量级的分类器都可以参与抽象级融合，因为从排序和度量信息可以得到决策信息。抽象级融合是应用最广泛、研究较早和较充分的一类融合方法，具有代表性的决策级融合方法有多数投票法、BKS方法、贝叶斯规则和证据理论方法等。

2. 排序级融合

在排序级融合中，分类器e把A（或子集$J \subseteq A$）中的标签按照某种规则排列成一个序列，排在首位的是第一选择。

单个分类器的输出是给定测试样本属于各类的可能性的一个排序列表。对于输入x，每个分类器$e_k(x)$产生一个子集$L_k \subseteq A$，且L_k中的标签排成一个序列。排序级融合是指利用事件$e_k(x) = L_k$，$k = 1, 2, \cdots, R$，构造一个集成的分类器E对x进行分类，输出一个确定的类别号，即$E(x) = j$，$j \in A \cup \{M+1\}$。

排序级融合要求所有的分类器输出排序级信息。任何输出度量级信息的分类器都可以参与排序级融合，因为从度量级信息可以生成一个排序列表。

3. 度量级融合

在度量级融合中，分类器e给每一个标签分配一个度量值，用以表示输入样本x属于相应类的程度。

在度量级融合中，单个分类器的输出是样本x属于相应类的程度。对于输入x，每个分类器$e_k(x)$产生一个度量向量$M_e(k) = [m_k(1), m_k(2), \cdots, m_k(M)]^{\mathrm{T}}$，其中$m_k(i)$表示$x$属于相应类$\omega_i$的程度。度量级融合就是利用事件$e_k(x) = M_e(k)$，$k = 1, 2, \cdots, R$，构造一个集成的分类器$E$对$x$进行分类，输出一个确定的类别号，即$E(x) = j$，$j \in A \cup \{M+1\}$。

在度量级融合中，要求所有的分类器输出度量级信息，并且具有相同的量级。有代表性的度量级融合方法包括贝叶斯方法、证据理论方法、模糊积分方法和神经网络方法等。

5.3.3.2　多分类器信息融合方法

1. 多数投票法

多数投票法是指在多个分类器中，选择得票最多的类别作为融合输出。

假设R个分类器$e_1, e_2, \cdots, e_R$对同一输入样本（模式）x进行分类，定义特征函数，即

$$T_k(x \in \omega_j) = \begin{cases} 1, & e_k(x) = j,\ j \in A \\ 0, & \text{其他} \end{cases} \tag{5.1}$$

最保守的投票规则是集成的分类器E把x划分到类ω_j，当且仅当所有R个分类器同时把x划分到类ω_j中，否则拒识x，即

$$E(x) = \begin{cases} j, & \exists j \in A,\ \min\limits_{k=1}^{R} T_k(x \in \omega_j) > 0 \\ M+1, & \text{其他} \end{cases} \tag{5.2}$$

多数投票准则为

$$E(x) = \begin{cases} j, & \sum\limits_{k=1}^{R} T_k(x \in \omega_j) = \max\limits_{i \in A} \sum\limits_{k=1}^{R} T_k(x \in \omega_j) > R/2 \\ M+1, & \text{其他} \end{cases} \tag{5.3}$$

更一般的形式为

$$E(x) = \begin{cases} j, & \sum\limits_{k=1}^{R} T_k(x \in \omega_j) = \max\limits_{i \in A} \sum\limits_{k=1}^{R} T_k(x \in \omega_j) > \alpha \cdot R \\ M+1, & \end{cases} \tag{5.4}$$

其中，$0 < a \leqslant 1$。若取$a = 1$，则式(5.4)变成式(5.2)；若取$\alpha = 0.5 + \varepsilon$，$\varepsilon$为很小的正数，则式(5.4)变成式(5.3)。

进一步修正式（5.4），得到新的多数投票准则为

$$E(x) = \begin{cases} j, & \sum_{k=1}^{R} T_k(x \in \omega_j) = \max_{i \in A} \sum_{k=1}^{R} T_k(x \in \omega_i), \\ & \sum_{k=1}^{R} T_k(x \in \omega_i) - \max_{i \in A-\{j\}} \sum_{k=1}^{R} T_k(x \in \omega_i) \geqslant \alpha \cdot R \\ M+1, & \text{其他} \end{cases} \tag{5.5}$$

式(5.5)不仅考虑了得票最多的类别号，而且考虑了得票第一多与第二多的票数差异。

多数投票法是决策级多分类器集成的最一般形式，其他决策级融合方法是投票准则的变形或改进。但是，多数投票准则存在如下问题。

(1) 所有的子分类器都是作为一票，没有考虑每个分类器性能的不同。

(2) 投票准则是基于抽象级信息的集成，在集成的过程中，很多有用信息被抛弃了。

(3) 对于每个子分类器，都需要确定一些门限值，如何确定这些值，并无理论指导。

2. BKS方法

BKS方法，又称为性能知识空间法，就是穷举各个分类器对训练样本集合的识别结果的各种组合，统计各种决策组合对应的样本，找出其中占主导地位的类别，作为多分类器融合的输出。

对于R个分类器$e_1, e_2, \cdots, e_R$的融合问题，一个性能知识空间（BKS）是一个R维空间，其中每一维代表一个分类器的判决结果，各个分类器的决策联合形成BKS的一个单元，称为局部单元。一个二维的BKS空间如表5.11所列，其中每一个分类器有$M+1$种可能的决策$\{1, 2, \cdots, M+1\}$。

表5.11 二维BKS空间

$e(1) \mid e(2)$	1	…	j	…	$M+1$
1	(1, 1)	…	$(1, j)$	…	$(1, M+1)$
…	…	…	…	…	…
i	$(i, 1)$	…	(i, j)	…	…
…	…	…	…	…	…
$M+1$	$(M+1, 1)$	…	$(M+1, j)$	…	$(M+1, M+1)$

对于给定的训练样本集合，利用R个分类器对所有的训练样本进行分类，并把各分类决策组合对应的样本归入相应的局部单元中。引入如下记号：

BKS的一个单元BKS（$e(1), e(2), \cdots, e(R)$）：第一个分类器给出了决策$e(1)$，第二个分类器给出了决策$e(2), \cdots$, 第R个分类器给出了决策$e(R)$等。

$N_{(e(1), e(2), \cdots, e(R))}(m)$：在BKS$(e(1), e(2), \cdots, e(R))$中属于类别$m$的训练样本总数。

$T_{(e(1), e(2), \cdots, e(\mathrm{R})}$：在BKS$(e(1), e(2), \cdots, e(R))$中的训练样本总数，即

$$T_{(e(1), e(2), \ldots, e(R))} = \sum_{m=1}^{M} N_{(e(1), e(2), \ldots, e(R))}(m) \tag{5.6}$$

$\Gamma_{(e(1), e(2), \cdots, e(R))}$：在BKS$(e(1), e(2), \cdots, e(R))$中占主导地位的类别，即$\Gamma_{(e(1), e(2), \cdots, e(R))} = j$，满足

$$N_{(e(1), e(2), \ldots, e(R)}(j) = \max_{1 \leqslant m \leqslant M} N_{(e(1), e(2), \ldots, e(R))}(m) \tag{5.7}$$

对于输入样本x，各分类器决策为$e(1), e(2), \cdots, e(R)$，则多分类器融合为

$$e(x)=\begin{cases}\Gamma_{(e(1),e(2),\ldots,e(R))}, & T_{(e(1),e(2),\ldots,e(R))}>0, \quad \dfrac{N_{(e(1),e(2),\ldots,e(k))}(\Gamma_{(e(1),e(2),\ldots,e(R))})}{T_{(e(1),e(2),\ldots,e(R))}}\geqslant\lambda \\ M+1, & \text{其他}\end{cases}$$

其中，门限$\lambda(0\leqslant\lambda\leqslant1)$用于控制最终决策的可靠性。

BKS方法是一种穷举的方法，克服了贝叶斯方法和D-S证据理论推理中分类器的独立性要求。BKS方法主要的不足体现在如下3个方面。

（1）需要大量存储空间。对于R个分类器和M类问题，BKS方法需要占用的空间是$O((M+1)^R)$，而贝叶斯理论只需$O(M\times(M+1)\times R)$。

（2）训练样本依赖性。BKS方法是建立在穷举分类组合基础上的，要求所选择训练样本分布要均匀，否则无法穷举到有些组合，特别是类别数增多时，无法穷举到的组合会更多。

（3）信息损失。BKS方法也是决策级信息融合，是一种投票准则的扩展，存在信息丢失问题。

5.3.4　JDL信息融合模型

信息融合可以从功能、结构和数学模型等几方面来研究和表示。功能模型从融合过程出发，描述信息融合系统的主要功能、数据库以及各组成部分之间的相互作用。结构模型描述信息融合系统的结构组成，包括软硬件配置、相关数据流、人机界面等。数学模型描述信息融合的算法和逻辑过程。

为了有效地组织和指导信息融合研究，美国三军政府组织——实验室理事联席会（Joint Directors of Laboratories，JDL）下设的C3技术委员会（TPC³）于20世纪80年代成立了信息融合专家组（DFS）。信息融合专家组制定了一个通用的信息融合功能模型——JDL模型。JDL模型能够促进系统管理人员、理论研究者、设计人员、评估人员相互之间的沟通和理解，使融合系统的设计、开发和实施过程得以高效顺利地进行，因此，已在越来越多的实际系统中采用。

JDL模型主要由数据源、数据预处理、一级处理（目标评估）、二级处理（态势估计）、三级处理（威胁估计）、四级处理（过程评估）、数据库管理系统以及人机交互操作等模块构成，其结构如图5.21所示。

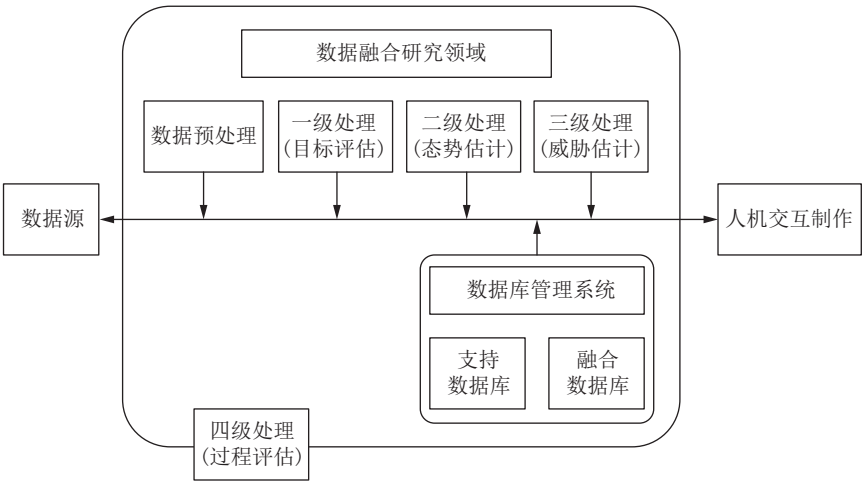


图5.21　JDL模型结构框图

JDL模型包括8个主要模块，其功能如下。

1. 数据源

数据源包括各类传感器及其相关数据（如人工情报、先验信息和环境参数等）。

2. 数据预处理

根据各观测平台、传感器以及数据源的特征，对来自不同传感器和数据库中的数据进行预处理，校正偏差，标准化输入数据以及提取关键信息，为融合中心提供重要的数据，并尽量减轻融合处理中心的计算负担。

3. 一级处理（目标评估）

目标评估就是对多传感器数据进行配准和关联，以获得目标位置、速度、属性、特征和身份等的精确可靠的估计，为态势分析和威胁估计提供目标信息。在目标评估之前，需要对目标进行检测，判断目标的有无，或区分不同的目标。

目标评估属于数值计算过程，包括数据关联、位置参数估计、运动参数估计、属性参数估计和身份估计。目标检测主要采用假设检验方法，包括二元假设和多元假设。

4. 二级处理（态势估计）

态势估计评价环境中各目标实体之间的关系、目标实体与环境之间的关系，以及这些关系随时间变化的趋势。态势估计包括态势抽象和态势评定。态势抽象是指构造一个综合的态势表示，解释实体之间的相互联系以及目标实体与环境之间的关系。态势评定是指获取事件态势的理解，其输入包括事件检测、状态估计以及为态势评定所产生的一组假设等，其输出是所考虑的各种假设的条件概率。

在军事上，态势估计是指对战场上战斗力量分配情况的评价过程，包括实体之间的相互关系、敌我双方兵力结构和使用特点等。

5. 三级处理（威胁估计）

威胁估计是指根据态势评估的结果，评估参与者的设想或行为的威胁、风险和影响。在军事上，威胁估计就是根据当前的态势，判断敌我双方的攻击能力和威胁程度等。在威胁估计中，除了考虑军事力量，还要考虑当前的政治环境和对敌策略等因素。因此，威胁估计是一个非常复杂的过程。

态势估计和威胁估计主要采用基于知识的推理方法，例如基于规则的黑板模型系统。

6. 四级处理（过程评估）

过程评估是指建立优化指标，对所有的信息融合处理进行监控与评价，实现多传感器及资源的最优配置与管理、信息的自适应获取与处理，以改善融合系统的性能。过程评估包括评价标准计算、传感器分配、资源优化等。

7. 数据库管理系统

数据库管理系统主要存储、检索、压缩和保护系统的数据和信息，主要包括传感器数据、支持数据库和一些中间处理结果。

8. 人机交互操作

人机交互操作提供人与计算机之间的交互功能，例如人机界面、操作员的指导与评价、多媒体功能等。

5.4　信息可视化技术

信息可视化指的是把抽象的、具有或不具有物理空间特征的信息转化成空间分布形式的图形图像，从而帮助用户理解或者发现其中隐藏的事物本质关系、形态和结构。信息可视化可以看成从信息到可视化形式再到人的感知系统的可调节的映射，它涉及人的视觉感知、计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互技术等多个领域。

战场环境可视化是信息可视化技术在军事上应用的典型例子。它可以为指挥员提供对战场的直观感受，大大提高了指挥员把握战场环境的准确度和速度，有利于指挥员迅速做出正确的决策。

本节将首先介绍人的视觉感知规律，然后介绍一些常用的信息可视化方法，最后介绍战场环境可视化技术。

5.4.1　视觉感知规律

人的视觉系统在认知周围环境的时候，首先会把所看到的世界分成背景和景物两部分。所谓背景就是不重要所以无须关注的部分；景物则是重要且需要关注的部分。例如，我们走在路上，离我们近的车辆就会成为景物，而道路、田野和远处的物体成为背景。

如果单纯考虑视觉效果，则越鲜艳、与背景差别越大的景物越吸引我们的注意力。在一片绿色的叶子中的一朵红色的花，会吸引我们的注意力。对于一大片既有绿叶又有红叶的树林，红叶更吸引我们的注意力。

从认知上来说，能够吸引我们的注意力的景物主要是有危害的和有价值的两类。所以，靠近我们的车辆会引起我们的注意，而且速度越快则我们越会关注它。看一棵果树的时候，果实往往比叶子更吸引我们的注意力，因为果实对我们更有价值。所以视觉规律和认知是密切相关的。

同时，由于视觉受到认知的影响，所以人的视觉不是简单地看到线条和小的结构，更多的是看到所关心的信息和整体的结构。这就形成了人的视觉的一些组织原则，可视化技术可以利用这些组织原则，其主要包括：贴近原则、相似原则、连续原则、闭合原则、共势原则、好图原则和经验原则。

1. 贴近原则

贴近原则是指如果很多小图形的位置非常贴近，则人们首先感受到的不是一个个的小图形，而是首先把它们看成一个整体。如图5.22所示，左边的方块距离较远，所以看上去就是一些零散的方块。而右边的图形则首先看到的是一个大大的“U”，然后才会注意构成这个“U”的各个小图形。

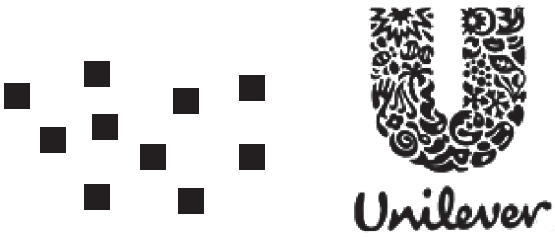


图5.22　贴近原则

2. 相似原则

相似原则是指如果有些小图形颜色相同，或者具有较明显的相同特征，就会被看成一个整体。如图5.23所示，左图的小圆点颜色深浅不同，浅色的是一类，深色的是一类。右图的小人也是一样，从上往下数第三排浅色的会被看成单独的一类，其他深色的则为一类。

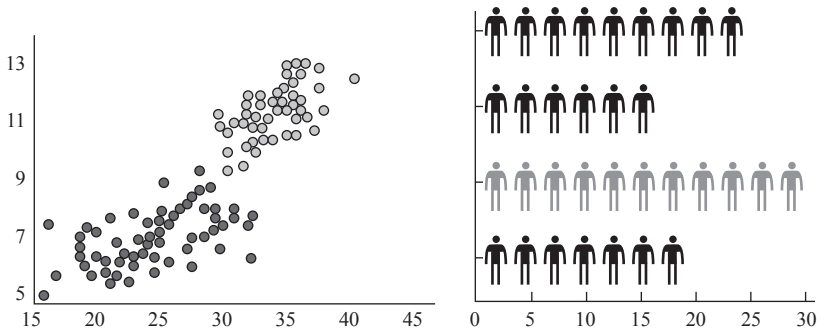


图5.23 相似原则

3. 连续原则

连续原则是指较近的两条线容易被看成一个连续的整体，按一定规律排列的比较接近的小图形也容易被看成一个整体。如图5.24所示，左图中虽然只是一些小圆点，但会被看成一个“*y*形”曲线，右图中断的两条曲线会被认为是同一条曲线，只是中间的部分信号被遮蔽了。

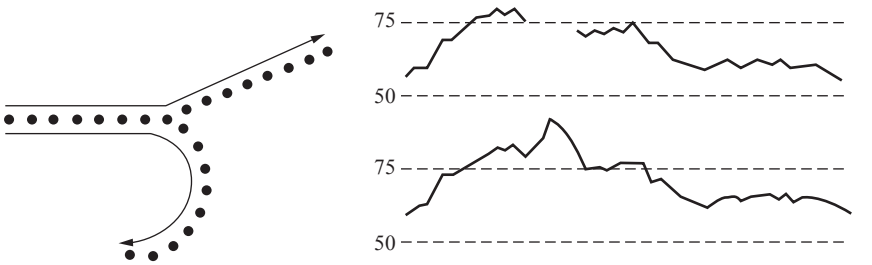


图5.24 连续原则

4. 闭合原则

闭合原则是指对于不连贯的有缺口的图形，会尽量使之闭合，构成完善的图形。如图5.25所示，我们一眼就看到了“IBM”，尽管它是由很多分离的线段构成的，但这些线段会自动被组织成一个闭合的图形。

图5.25 闭合原则

5. 共势原则

共势原则是指对于一些具有相同运动方向和速度的小图形，会构成一个整体。由于人的视觉经验，具有相同运动趋势的小图形，或者具有相同姿态的小图形，也会构成一个整体。如图5.26所示，左图的中间一行字母构成一句完整的话，尽管周围有很多其他形态的字母，但由于这一行字母的形态相同，所以被看成一个整体。而右图“2009”也会被看成一个整体，因为在图中，这几个数字具有相同的大小和姿态。

6. 好图原则

好图原则是指如果图形是残缺不全的，会被自动补充残缺的部分，从而成为一张好图。如图5.27所示，右边的图形显然不完整，但我们看上去的时候，仍然会把它看成奥运五环，人的认知系统根据经验，自动补充了残缺的内容，构成一个好图，作为认知的结果。

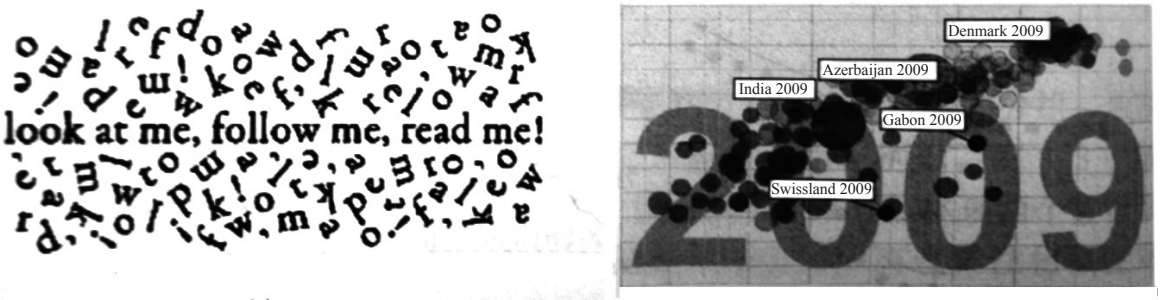


图5.26　共势原则

图5.27　好图原则

7. 经验原则

经验原则是指人会根据经验和当前的环境，把图形认知为相应的结果。如图5.28所示，左侧图形为“A B C”，右侧图形为“12 13 14”。仔细观察图形，会发现“B”和“13”其实是相同的图形。由于我们运用了经验，所以当这个图形和A、C在一起的时候就会被看成“B”，而和12、14在一起的时候，则会被看成“13”。

图5.28　经验原则

5.4.2　视觉通道特点

视觉通道指的是图形元素的特性，比如图形元素的位置、尺寸、颜色、亮度、色调、饱和度、配色方案、透明度、方向、形状、纹理、运动等。这些视觉通道可以分成两类：一类是定性的视觉通道，例如形状、颜色、亮度等；另一类是定量的视觉通道，例如直线的长度、区域的面积等。有些视觉通道具有定性、定量两种特征，如位置，既可以用来表现分类，也可以用来表达数量。

不同的视觉通道具有不同的表现力。在设计可视化图形时，对于采用什么样的视觉通道，应当从准确性、可辨认性、可分离性、视觉突出等方面来考虑和比较。准确性，是指视觉通道是否能够准确表达数据之间的变化；可辨认性，是指同一个视觉通道能够编码的分类个数，即可辨识的分类个数上限；可分离性，是指不同视觉通道的编码对象放置到一起，是否容易分辨；视觉突出，是指视觉通道是否适合突出表达重要的信息。

对于视觉通道适合表达的信息，所具有的规律如图5.29所示。

不同的视觉通道，具有不同的准确性，如图5.30所示。

在设计可视化方案时，就是针对所要表达的信息，选择合适的视觉通道，经过多个视觉通道的适当组合，通过显示界面表达，达到满意的用户认知效果。

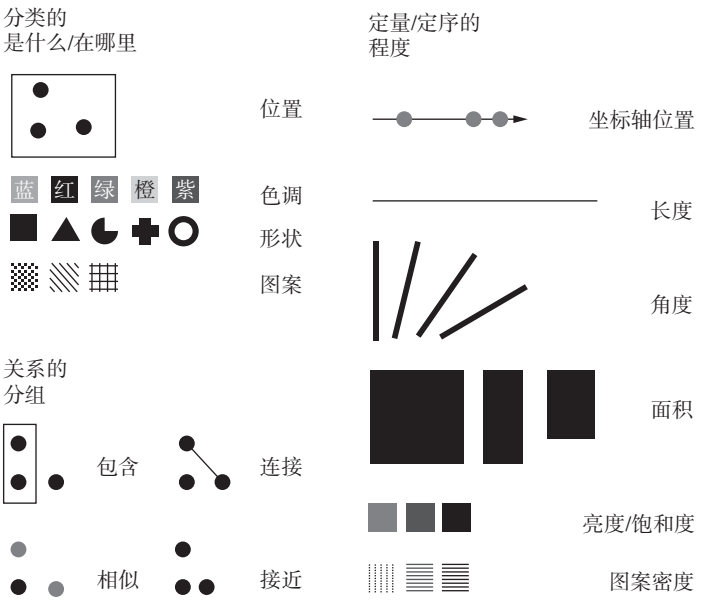


图5.29 视觉通道的分类

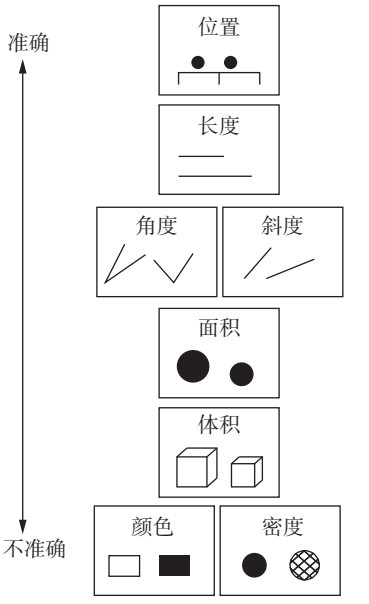


图5.30 视觉通道表达信息的准确性

5.4.3 常用可视化方法

信息可视化方法有很多种，这里介绍基于几何投影的技术、基于图标的技术和基于层次的技术。

5.4.3.1 基于几何投影的技术

基于几何投影的技术的目的是在多维数据集中找到“有意义的”投影，以几何画法或几何投影的方式来表示数据库中的数据，主要有散点图、地形图、投影寻踪、平行坐标图等。

1. 平行坐标

平行坐标技术由Inselberg和Dimsdale提出，用于表示多维信息或数据，将多维信息映射到一组平行的等距离坐标轴上。基本思想是将n维数据属性空间通过n条等距离的平行轴映射到二维平面上，每一条轴线代表一个属性维，轴线上的取值范围对应属性从最小值到最大值均匀分布。这样，每一个数据项都可以根据其属性值用一条折线段在n条平行轴上表示出来，相似的对象就具有相似的折线走向趋势。

例如，假设数据库中有关于主战坦克的数据，如表5.12所示。

表5.12 几种主战坦克的数据

坦克型号	乘员数（人）	火炮口径（mm）	战斗全重（吨）	最大速度（km/h）	最大里程（km）
拉姆塞斯2	4	105	43	69	530
MB-3塔穆伊奥	4	90	30	67	550
EE-T1	4	105	35	70	570
AMX-30B	4	105	36	65	500
AMX勒克莱尔	3	120	53	71	550

如果将上表数据用平行坐标表示，则规则为：这个表的数据为5维，用5条纵坐标分别表示，根据每维的最大值、最小值设置纵坐标的刻度，以x轴处的纵坐标为各维的最小值，尽量使差别最大化，如图5.31所示。

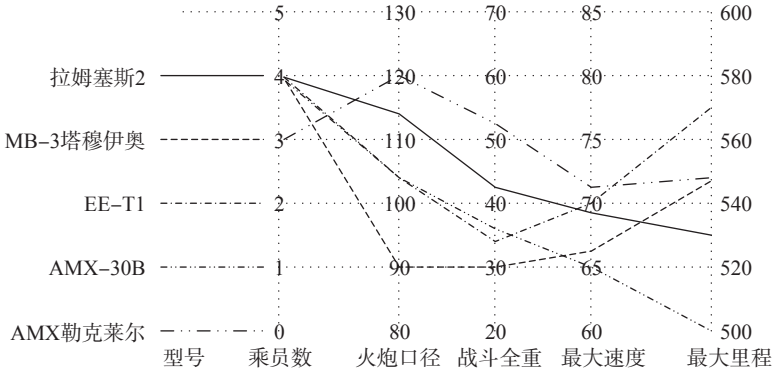


图5.31　几种主战坦克的数据

从图5.31可以看出，坦克的乘员数明显集中于4人，说明大多是4人，这和坦克作战、维护、保养等的需求有关。坦克的速度则在60~70 km之间，火炮口径和战斗全重具有一定的联系，战斗全重小的坦克所配备的火炮口径也小。用类似的方法，绘制出多种数据的平行坐标图以后，有可能会发现其中存在着某些我们尚未发现的规律。

2. 散点图

散点图是将多维数据以平面或空间中的点来表示，最常用的是二维数据在笛卡儿坐标系内表示的情况，称为直角散点图或XY散点图。有时为了更好地描述多维数据的变化趋势，用直线或平滑曲线将各数据点连接起来，称为折线散点图和平滑线散点图。

例如，在前面几种主战坦克的数据中，可以单独取出火炮口径和战斗全重两维数据绘制成散点图（见图5.32）。由图可以看出，火炮口径和战斗全重接近于线性关系。

每个散点图只能表示二维数据，如果要表示多维数据，则可以把散点图组合起来，构成散点图矩阵。

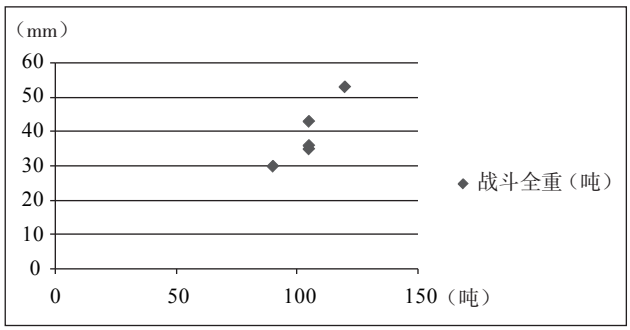


图5.32　火炮口径和战斗全重的散点图

3. 雷达图

雷达图又称蛛网图，由于图形与导航雷达显示屏上的图形十分相似而得名。其主要特点是直观，通常用来进行定性评价，是目前广泛应用的多维数据可视化方法。

绘制雷达图的步骤如下。

（1）设要分析的数据共有n个变量，先画一个圆，用n个点把圆周等分成n个部分；

（2）将圆心和n个点连接起来，即可得到n个辐射状的半径，这n个半径就作为n个变量的坐标轴；

（3）为划分刻度方便，在标记坐标轴前需要对原始数据进行线性变换，使得数据落在给定区间[0, r]内，其中r为雷达图显示界面的半径。根据下面的公式，对于第i个多维数据中第j个分量作进行线性变换：

$$y_{ij} = \frac{x_{ij} - x_{\min,j}}{R_j}$$

其中，$x_{\min,j}=\min\limits_{1\leqslant i\leqslant n}x_{ij}$，$x_{\max,j}=\max\limits_{1\leqslant i\leqslant n}x_{ij}$，$R_j=x_{\max,j}-x_{\min,j}$。

将n维数据的各个维的归一化数值刻在对应的坐标轴上，依次连接起来得到一个n边形，即得到用平面表示的n维数据的雷达图。

当然，也可以不进行归一化，而是将相应的数据进行线性变换。如何变换数据，取决于数据的含义和看起来是否方便，是否容易看出规律。

对于前面的几种主战坦克数据，由于乘员数和其他属性对坦克性能的反应规律不同，因此去掉了其中的乘员数。剩余的属性基本上均为值越大则坦克威力越大，因此在将其绘制成雷达图时，圈住的面积越大，则说明坦克威力越大。绘制的雷达图如图5.33所示。注意，为了看起来方便，除最大里程以外，对其他数据分别乘了一个系数。

当要分析的多维数据个数较少时，可以在同一个雷达图中将它们表示出来；当维数较大时，为使图形清晰，每张图形可以只画少数几个样本数据；或者根据数据的相关性将它们分组，同一组的用同一个雷达图表示，其中不同的多维数据可用不同颜色的多边形来区分。

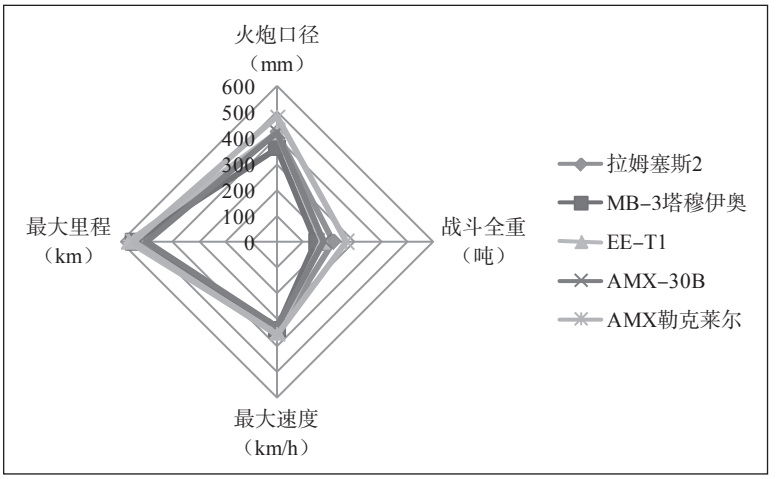


图5.33　几种主战坦克性能的雷达图

为了获得较好的效果，在雷达图中，适当分配变量的坐标轴并选取合适的尺度，是十分重要的。例如，把要进行对比的指标分别放在其坐标轴左和右或正上方和正下方，以便根据图形偏左、偏右，或偏上、偏下进行对比和分析。还应注意，这里的坐标轴只有正半轴，因此只能表示非负数据，若有负数据则需要变换成非负数据。

4. 气泡图

气泡图与XY散点图类似，但是它们对成组的三个数值而非两个数值进行比较。第三个数值确定气泡数据点的大小。从这个意义上讲，气泡图可以表达三维信息。要使气泡图清晰可见，要求作为坐标轴的两维是离散的，不能是连续的。

例如表5.13所示的数据，可以用气泡图表示（见图5.34）。

表5.13　待维修武器数量

待维修武器数量	步枪	机枪	火炮	装甲车
一连	5	3	2	2
二连	3	0	1	0
三连	2	1	1	1

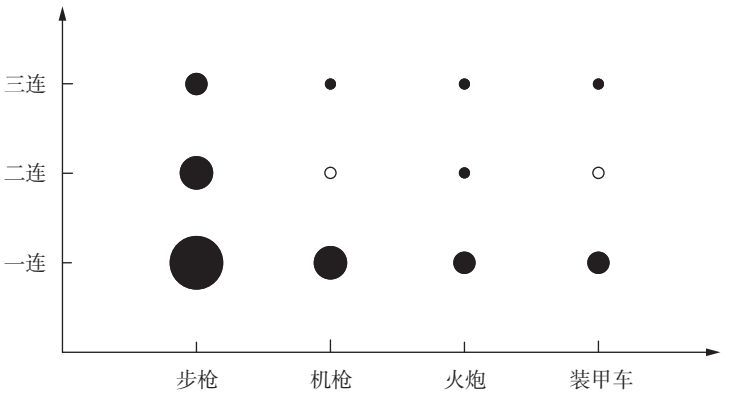


图5.34　气泡图示例

在气泡图中，要根据数据的含义设置气泡的大小，以使人的认知和真实的数据大小关系相吻合。根据需要，可以在气泡的旁边标出数值的大小。在一个气泡图中，也可以表示多种数据，例如还可以把武器装备的损坏率用气泡显示在图中。

5. 星座图

星座图经常用于聚类分析。星座图的基本形式是一个半圆，如图5.35所示，把数据按照一定的方式进行变换，使每个数据点对应半圆中的一点，根据该点的绘制路径或位置进行聚类分析。

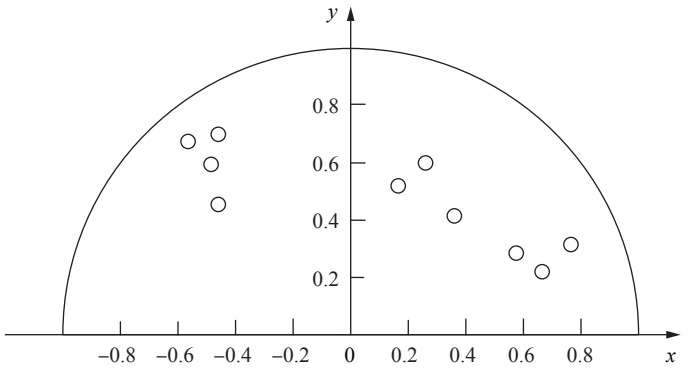


图5.35　星座图示例

数据的变换方法很多，根据实际数据的物理含义，可以选择或设计合适的变换公式。下边给出一个示例。星座图聚类法步骤如下。

（1）对原始数据进行极差变换，并使变换后的数值落于[0°, 180°]的闭合区间内：

$$\phi_{ij}=\frac{X_{ij}-X_{j\min}}{X_{j\max}-X_{j\min}}\times 180$$

其中，ϕ_{ij}为变换后的数据；X_{ij}为第i行第j列的数据；$X_{j\max}$和$X_{j\min}$分别为第j列数据的最大值和最小值。

（2）对第i行数据根据其物理含义中对结果的影响程度赋予一个权重W_{ij}，使$0 < W_{ij} < 1$，$\sum_{j=1}^{n} W_{ij}=1$。

(3) 计算笛卡儿坐标值x_{ij}和y_{ij}:

$$x_{ij}=\sum_{j=1}^{n}W_{ij}\cos\varphi_{ij}$$

$$y_{ij}=\sum_{j=1}^{n}W_{ij}\sin\varphi_{ij}$$

(4) 绘制星座图。以(0, 0)为坐标原点作x轴、y轴，根据以上计算各点的x_i和y_i的值，确定每个数据点的位置，性质相似和相近的点便聚在一起，形成一个“星座”。由于数据进行了极差变换，且$\sum_{j=1}^{n}W_{ij}=1$，故所有的点均落在以(0, 0)为圆心、半径为1的半圆中。

为了计算简便，没有给出星座图中各点对应的数据，只作为示意。从星座图中各点的分布情况可以看出其关系的远近，位置相近的数据点可以归为一类，所以星座图常用于聚类分析。要用好星座图，主要工作就是设计好数据点到图形位置的映射，以使它们的位置和它们之间的关联关系相对应。

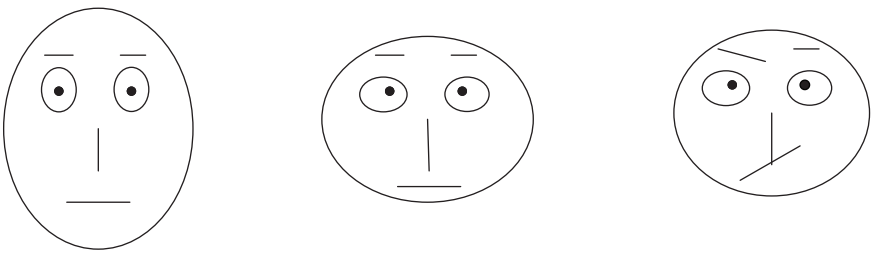
5.4.3.2 基于图标技术

基于图标技术的基本思想是将每一个多维数据项映射成一个图标，即用一个简单图标的各个部分来表示n维数据的属性。基于图标的可视化技术包括脸谱图、枝形图等。

1. 脸谱图

Chernoff于1973年提出用于显示多元数据的脸谱图。目前脸谱图已经广泛应用于统计图表示和数据可视化等领域。

脸谱图的基本思想是将数据项的两个维度映射成两个用于显示的坐标维，而剩下的维度则映射成一张脸的各个器官——鼻子、嘴巴、眼睛的现状以及脸部本身的形状。如图5.36所示。图中(a)为正常图形，(b)和(c)都是某些参数出现偏差时的图形。显然，从图中可以很直观地看出数据有没有问题，什么数据有问题。



(a) (b) (c)

图5.36 脸谱图

要用好脸谱图，非常重要的就是设计好数据和脸部特征的对应关系，应当达到的效果如下：

(1) 当各数据都处于较好的状态时，脸谱形状正好匀称、漂亮；

(2) 当有的数据处于不好的状态时，脸谱会发生相应的变形，例如某些器官变大、变歪等；

(3) 当数据代表的状态非常不好时，脸谱应当处于哭脸的形态；

(4) 当数据代表的状态有危险时，脸谱应当显示惊恐、呼喊的形态；

(5) 当数据代表的状态说明某些工作有失误时，脸谱应当处于惊讶、生气、愤怒等形态等。

总而言之，脸谱图的形态一定要和数据代表的状态相对应，才能达到直观显示的效果。

2. 枝形图

枝形图方法选取多维属性中的两种属性作为基本的*Oxy*平面轴，在此平面上利用小树枝的长度或角度的不同，表示出其他属性值的变化。例如，图5.37所示的两个数据点，它们左边的两个属性具有相同的数据值，而右边的两个属性的数据值则不相同。

枝形图应用的关键和脸谱图类似，也是如何把属性值和图形的状态对应，从而表示出数据真实的含义，即把数据的含义和人对图形的认知规律统一起来。

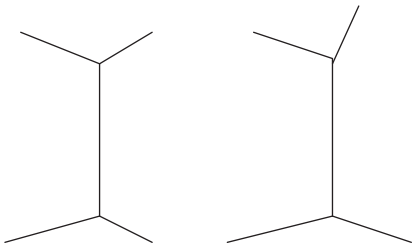
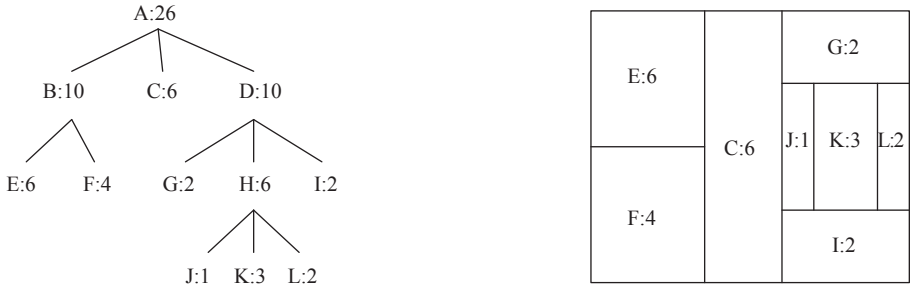


图5.37　两个数据点的枝形图

5.4.3.3　基于层次的技术

基于层次的可视化技术主要针对数据库系统中具有层次结构的数据信息，如人事组织、文件目录、人口调查数据等。它的基本思想是将*n*维数据空间划分为若干子空间，对那些子空间以层次结构的方式组织并以图形表示出来。树形图是其中的一种代表技术。如图5.38所示，图(a)中的树形结构数据在图(b)中以树形图表示，图中每一个节点都有一个名称和数值大小，父节点是各子节点大小的总和。例如节点“A:26”表示：节点名字为A，大小为26。

树形图是根据数据的层次结构将屏幕空间划分成一个个矩形子空间，子空间大小由节点大小决定。树形图层次则依据由根节点到叶节点的顺序，水平和垂直地依次转换，开始将空间水平划分，下一层将得到的子空间垂直划分，再下一层又进行水平划分，依此类推。对于每一个划分的矩形可以进行相应的颜色匹配或必要的说明。



(a)树形结构数据　　(b)树形图

图5.38　树形结构数据与树形图

在图5.38中，带有子节点的父节点A、B、D和H没有在图上标出其名称，在实际使用时，可根据需要，采用适当的方式，例如采用加宽的带有特殊纹理的边框，更醒目地标出其边界并加上标题和说明。

5.4.4　战场环境可视化

战场环境是战场及其周围对作战有影响的情况和条件，包括自然条件和人文条件，如地理、气象、人口、交通、物资、场站建设等。一般来说，战场环境包含战场物质环境和战场信息环

境两个基本要素。这两大要素又由次一级的子要素组成，如地理环境、气象环境、电磁环境、核生化环境等。

由于在战场环境中各类子要素互相影响、渗透，对它们的界定和分类有许多不同的观点，没有一个明确的标准，为便于对其进行可视化方面的研究，可以把战场环境的构成要素分成两大基本要素及七个子要素，如图5.39所示。

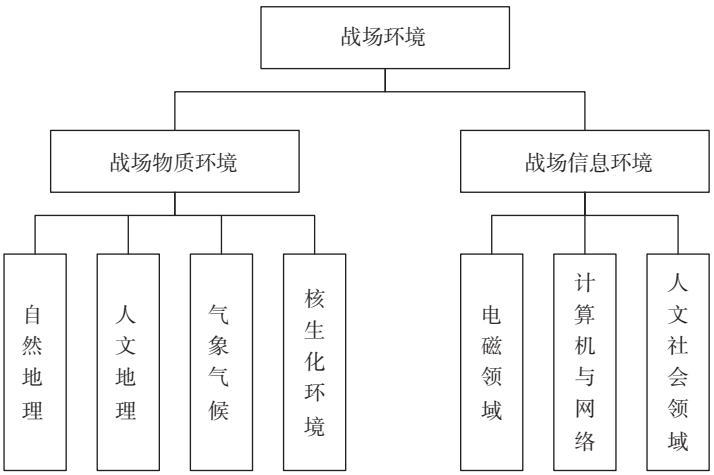


图5.39 战场环境构成要素

通俗地说，战场环境的可视化就是运用可视化技术来表达战场环境的各组成要素。其目的是构成直观鲜明、信息丰富的以电子形式显示的战场详图或虚拟战场，方便指挥员感知态势，定下决心。战场环境的可视化不仅应用于战场指挥，对于训练、战术评估、论证系统也有重要的意义。

战场环境可视化通常应提供如下一些功能。

(1) 空间测绘数据可视化功能。包括太空、地形和海洋在内的测绘数据的可视化，实现真实视景与信息符号化表达的结合，既直观，又有丰富的信息。

(2) 城市环境可视化功能。城市作战非常复杂，通过把城市的道路、建筑、桥梁、水体，以及政府和企业等各种组织分布信息可视化，使指战员了解和掌握城市环境，使军事行动准确高效，往往可以达到“以弱胜强、以少胜多”的奇效。

(3) 战场气象环境可视化功能。气象环境是影响作战的重要自然因素之一，通过把天气、空气温度、空气湿度、气流、水流、水温、海水盐度等各种气象指标可视化，可以满足指挥员对掌握战场气象环境信息的需求。

(4) 战场电磁环境可视化。通过把电磁信号的种类、强度、信号源、方向、频率以及各种装备的抗电磁干扰能力等信息以可视化形式展现，可以使指战员直观地了解电磁环境状态，更合理地配置装备和部队，减少电磁干扰，提高电磁对抗能力。

(5) 多比例尺、多分辨率的战场地理环境可视化功能。按照不同比例尺、不同分辨率描述战场环境，不仅满足了指挥员对敌我的位置、态势、部署、运动、损耗及战果情况的了解，同时也满足作战分队对战场的详细信息的需求。

战场环境可视化最常见的应用就是作战态势图。作战态势图一般是在地理信息基础上，叠加显示指挥员关心的各种战场信息，这些战场信息就是态势要素，大致可分为以下五类：兵力

部署与作战能力类、重要动态目标类、对抗措施类、战场环境类，以及政治、经济和社会环境类（见图5.40）。

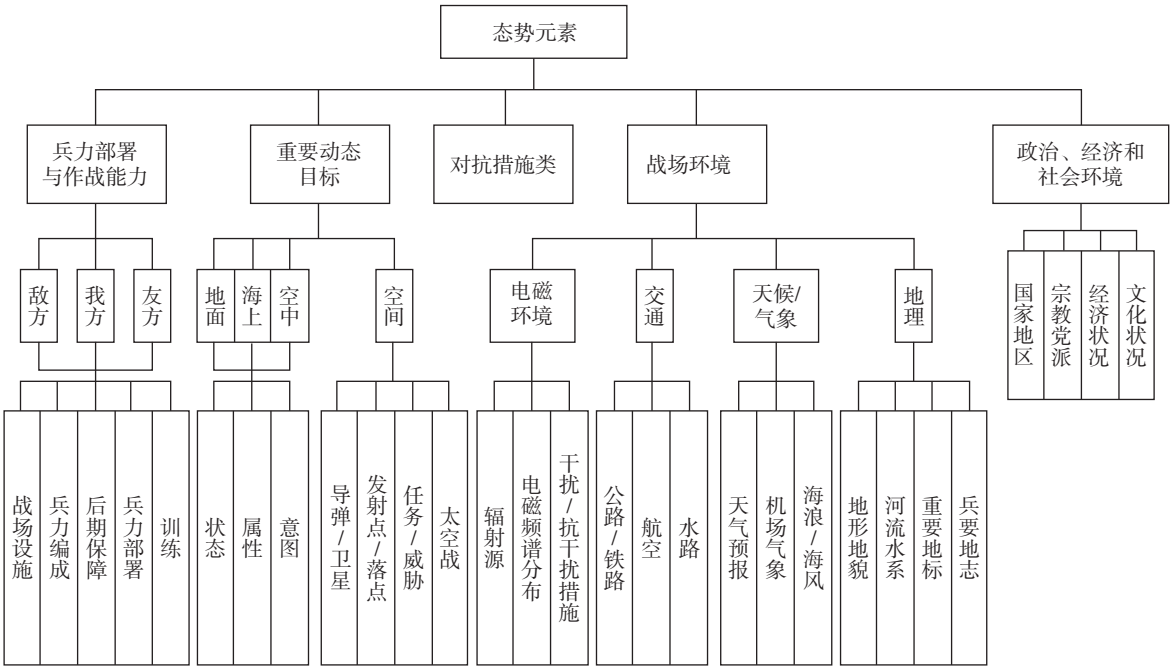


图5.40　作战态势图的态势要素

在电子化地理信息系统出现之前，一般是在大比例尺地图上进行各种信息的标注，构成态势图，但手工操作很不方便。现在，有了电子化的地理信息系统，部队配发大比例尺的电子地图信息并显示在电子屏幕上，就可以通过人机交互，在电子地图背景上标注各种战场信息。由于实现了动态刷新，可以不断显示出战场的最新态势，大大提高了指挥员的态势感知能力。

近年对作战态势图的研究主要集中于通用作战态势图，即通过强大可靠的网络能力和信息处理能力，集中全系统的信息，为各级指挥员和战斗员提供统一的态势图。通过让所有人看到同样的战场态势图，提高信息共享能力，提高各作战单元之间的协同能力。

后来，在实际应用中发现，不同层次的作战人员所关心的战场范围、战术对象、态势要素及内容、信息精度与时效等都存在差异，同一层次的作战人员由于所处职能部门或者分管业务不同，所关心的往往是各自业务职责有关的战场态势要素，对全局态势只需要粗粒度地了解。同时，在制定协同计划时，由于协同诸方有各自的态势信息源，在共用的态势图上，只需要保证共同关心的态势元素相同即可，其他态势元素并不需要完全一致。所以，提出了更加切合实际应用的用户定义作战态势图的概念，使用户能够根据需要主动提取态势信息，而不是等待系统不加选择地分发相同的态势画面。系统发展到“可以讨论和组合不同视角观点的协作环境”，通用态势图从态势的可视化显示发展到相关态势数据与信息的共享。

当然，不管怎么变化，作战态势图的根本目的并未发生变化，即为了实现所有指挥员和战斗员对态势的一致理解。

思考题

1. 什么是信息融合？信息融合主要分成哪几类？它们的原理分别是什么？
2. 什么是模式？什么是模式识别？
3. 什么是最近邻法？什么是*k*-近邻法？ *k*-近邻法比最近邻法有什么优点？
4. 在一个二维平面上的已知一些点，分为A、B、C三类，A类中有(2, 1)、(1.5, 1.5)、(2, 2)，B类中有(3, 2)、(3, 2.5)、(3.5, 2)、(3.5, 2.5)、(4, 2)、(4, 2.5)，C类中有(5, 1)、(5, 2)、(5, 3)。现有一个点x=(2.4, 2)，分别用最近邻法和*k*-近邻法(k = 3)对其进行分类。
5. 根据5.1.3节中的购物篮数据，请计算{面包}⇒{黄油}的支持度和可信度。如果设最小支持度为0.3，最小可信度为0.8，那么{面包}⇒{黄油}这个规则成立吗？
6. 什么是一元线性回归和多元线性回归？
7. 非线性问题有可能用线性回归方法解决吗？请举例说明。
8. 已知10个学生的成绩：51、62、77、49、65、66、53、71、65、61。请进行聚类分析，把他们分成三个类别。
9. 试设想数据挖掘技术在军事中的应用。
10. 什么是可视化技术？试设想在作战和保障中如何应用可视化技术。

第6章　信息服务技术

本章将介绍信息服务相关技术，即信息资源的组织、检索、导航和推荐等。本章所说信息的内涵更侧重于经过加工处理后的信息，即知识或情报。“云平台”技术是信息技术领域刚刚发展起来的新亮点，对于信息服务有良好的应用前景，本章最后一节将单独予以介绍。

6.1　信息资源

从某种意义上讲，人类社会活动的过程也是交流和利用信息资源的过程。信息资源是指经过一定加工处理，具有一定价值的信息集合。从应用的角度讲，信息资源大致可分为两类，即情报信息资源和网络信息资源。

6.1.1　情报信息资源

虽然关于情报的科学概念一直众说纷纭，但图书馆学都是其中的核心。一般而言，图书馆资源主要指的是文献资源，其主要包括以下几种类型。

1. 图书

图书主要指的是以印刷方式单本刊行的出版物，包括汇编本、多卷书、丛书等。它是人们为系统地传播知识或经验而写成并出版的文献。由于图书往往以原始记录为素材，对某一领域进行系统阐述，或对现有成果、技术和经验进行归纳和概括而成，因此内容比较成熟、全面、系统、可靠且具有一定的新颖性。由于其编著和出版的周期较长，因而其内容一般缺乏最新的研究成果，但是对于了解和掌握某一学科的系统知识，有重要参考价值。

2. 报纸

报纸是以刊载新闻和评论为主的定期出版物。它具有固定名称，多数每日出版，也有隔日或每周出版的。随着现代社会生活的发展和读者对信息需求的多样化，除了以传播新闻和评论为主的报纸外，还有以传播知识、提供娱乐或生活服务为内容的报纸。报纸因其具有时事性、普及性、大众性、服务性，以及传递信息迅速且信息量大的特点，受到广大读者的喜爱，成为一种十分重要的信息来源。

3. 期刊

期刊有固定的名称，定期或不定期发行，用卷、期或年、月顺序编号，成册连续出版。期刊由于出版的周期短，能及时反映新理论、新技术和新方法，且知识新颖，起到了迅速传播新研究成果的作用。

4. 特种文献

(1) 科技报告。科技报告是科研人员进行项目研究的实际记录，包括阶段报告、成果报告和总结报告等，多以其编辑出版机构的名称作为总的固定名称，并标有连续编号，每一报告为一项专题资料，自成一篇（册），不定期出版。科技报告研究的多是重大的课题，涉及基础理论和应用技术研究的各个重要领域，往往代表一个国家或专业领域的最高研究水平，因而很受广大科研人员和生产单位的重视。

(2) 政府出版物。政府出版物是各国政府及其所属部门公开发表、出版的各类文献的总称。又分为行政性文件和科技文献。

(3) 会议文献。会议文献是指在各种国内外学术会议上宣读和交流的论文、报告及其他有关文献。

(4) 专利文献。专利文献是记载有关发明创造信息的文献，包括专利说明书、专利申请书、专利公报、专利文摘、专利索引和专利分类表等。

(5) 技术标准。技术标准是标准化组织或有关机构对工农业产品和零部件的质量、规格、生产过程和检验方法等所做的技术规定。

(6) 学位论文。学位论文是高等学校和科研单位中的学生为获取学士、硕士或博士学位而撰写提交的学术论文。

(7) 产品文献。产品文献是企业为宣传自己的产品而编印的有关资料，包括产品样本、产品说明书、产品目录和产品广告等。

5. 缩微文献

缩微文献又称缩微复制品，采用摄影的方法，将文献缩小复制在感光材料上，然后借助阅读设备进行阅读。

6. 机读文献

机读文献是将文字、声音、图形、图像等信息以数字代码方式存储在磁、光、电等介质上，通过计算机或类似功能的设备进行阅读和使用的文献。

7. 视听文献

视听文献是以电磁材料为载体，以电磁波为信息符号，将声音、图像和文字记录下来的一种动态型文献，它必须通过视听设备才能阅读。

6.1.2 网络信息资源

目前，学术界对“网络信息资源”的含义说法不一，较有代表性的观点包括以下几种。

(1) 网络信息资源是一切投入互联网络的数字化信息资源的统称。它们分布在不同的网络节点上，可以利用现代信息技术进行制作、加工、传输、转换和二次开发。与传统的信息资源一样，网络信息资源涉及人类生产、生活、娱乐及其他社会生活的各个方面，是随人类社会实践的发展而不断累积起来的。

(2) 网络信息资源是指以数字形式记录，以多媒体形式表达，存储在网络计算机磁介质、光介质及各类通信介质上的信息集合。

(3) 网络信息资源是将文字、图像、声音、动画等多种形式的信息，以数字化的形式存储，并借助计算机与网络通信设备发布、收集、组织、存储、传递、检索和利用的信息资源。

(4) 网络信息资源是“通过计算机网络可以利用的各种信息资源的总和”。

从根本上讲，“网络信息资源”的概念是随着因特网的发展和网上信息资源的增长，以及由此而引发的对网上信息资源管理和开发利用的社会需求而产生的。我们可以认为，“网络信息资源”是通过因特网可利用的各种信息资源的总和，其在本质上属于“信息资源”的范畴。因此，与信息资源一样，网络信息资源也可以从广义和狭义两个层次来理解。狭义的网络信息资源是“以数字化形式记录的，以多媒体形式表达的，存储在网络计算机磁介质、光介质以及各类通信介质上的，并通过计算机网络进行传递的信息内容的集合”。广义的网络信息资源是网络信息活动中所有要素的总和，包括与网络相关的信息内容、信息网络、信息人才、信息系统、信息技术等资源。

事实上，网络信息资源是一种与传统资源有很大差别的新型数字化资源。在网络环境下，信息以计算机可识别的方式存储于网络的某一节点上，并且可以在任何需要的时候通过四通八达的全球互联网络传向任一合法的网络终端用户。

与传统的信息资源相比，网络信息资源具有以下特点。

(1) 内容丰富，形式多样。因特网已经成为当代信息存储与传播的主要媒介之一，也是一个巨大的信息资源库。从对网络信息资源类型的阐述和研究中可以看出，其内容包罗万象，覆盖了不同学科、不同领域、不同地域、不同语言的信息资源。在形式上包括了文本、图像、声音、软件、数据库等，堪称多媒体、多语种、多类型信息的混合体。

(2) 分布开放、关联度强。网络信息资源在分布上极其广泛，打破了信息资源分布的时空阻碍。网络特有的超文本链接方式，使得内容之间具有很强的关联程度。

(3) 传输速度快。因特网提供了覆盖全球的高速的信息传输渠道，通过互联网络实现了网络信息资源的即时传递，使信息资源更加快捷地应用于各个领域。

(4) 共享程度高。在网络环境下，时间和空间范围得到了最大程度的延伸和扩展。网络信息资源可以同时提供给多个用户使用，而且在使用过程中信息内容不会损失。

(5) 使用成本低、方便快捷。在因特网上，大部分信息资源都可以免费使用，用户只需支付网络使用费用。虽然还有一些有偿的网络信息资源，但是与其他形式的信息资源相比，网络信息资源在满足用户信息需求的情况下，节省了大量的人力和时间成本。

(6) 信息的分布和构成缺乏组织，结构混乱。网络信息资源中存在大量未经组织和规范的信息，使得网络信息资源的分布具有很大的随意性，无论是信息的存储地点还是信息形式上的分类都比较混乱，不便于用户使用。

(7) 质量良莠不齐。因特网存在着巨大的开放性，用户在存储和发布信息时有很大的自由度，从而导致了大量冗余、粗制滥造甚至虚假信息的存在。无用信息与有用信息混杂在一起，给用户带来诸多不便。

(8) 交互性强。在因特网上，信息发布具有很大的自由性和任意性。用户既是信息资源的使用者，同时也是信息资源的生产者。用户可以实时地利用和提供网络信息资源，在这个过程中并不受限制。

(9) 信息关联度强，检索快捷。网络信息资源利用超文本链接，构成了立体网状文献链，把不同国家、不同地区、各种服务器、各种网页、各种不同文献通过节点链接起来，关联度随之增强。通过专用搜索引擎及检索系统，可使信息检索变得方便快捷。

目前，随着计算机网络的逐步普及、数字文本复制的便利和自由发表的实现，使得网络信息资源数量急剧增加。但是，社会信息量的增长并不意味着用户获取的信息量的增长，恰恰相

反，无序的信息资源不仅无助于信息资源的使用，反而会加剧信息增长与使用的矛盾。人们生动地称这种情况是“信息超载，知识缺乏”。越来越多的人认识到，“原始信息本身并不能产生价值”，只有将其有效地组织起来，按特定的需要集中和揭示，才能产生价值。

信息服务技术，就是以手工和计算机方式对信息进行加工、处理，使其有序化，便于快速查找并传递给有特定需求的用户，这是有效利用信息资源的基础。为了向用户提供良好的信息服务，需要对信息资源进行有效的组织，并提供方便的信息检索方法，有的还借助信息导航和信息推荐技术，最大程度地方便用户准确地获取和利用所需的信息。

6.2 信息组织

信息组织是对信息的外部特征和内容特征进行揭示和有序化。信息的外部特征一般是指其载体的物理形态、题名、责任者、出版项等。对信息外部特征进行的组织与记录称为信息描述，根据特定的信息管理规则和技术标准，将存在于某一物理载体上的信息的外在特征进行选择和记录。信息的内容特征一般是指其学科专业归属或主题概念。对信息内容特征进行的组织加工称为信息标引，是在分析信息主题内容的基础上，根据特定的标引规则与语义工具，赋予信息内容一定的标识，以便将信息记录组成概念标识系统。

6.2.1 信息描述

信息描述主要指对信息外部形式特征的描述，其目的是以描述记录为中介，对信息资源进行识别确认，组织定位并组织检索系统，以便选择利用。

信息描述要根据不同类型信息资源的特点和检索需求，遵循相应的著录规则进行。图书文献的著录项目一般为九大项：题名和责任者项、版本项、文献特殊细节项、出版发行项、载体形态项、丛编项、附注项、文献标准编号及有关记载项、提要项。各大项有的又包括若干小项。档案著录项目一般为七大项：题名与责任说明项、稿本与文种项、密级与保管期限项、时间项、载体形态项、附注与提要项、排检与编号项。各大项有的也包括若干小项。特殊类型的信息资源可设置特定的著录项目，如非书资料可著录项包括数量、材质、播放时间、制式、色别等。此外，一个完整的记录一般还包括标识项。传统文献款目的标识项通常由题名、责任者、分类号、主题词等组成。

信息描述的结果是获得描述记录，也就是元数据，以此来组织检索系统，因此信息描述必须客观地反映信息资源的特征，做到准确、规范、完备。准确，即对信息资源的描述应真实反映其各种特征；规范，即对信息资源的描述应严格遵守相应的描述规范，采用规范的文档、规范的格式及规范的语言；完备，即对信息资源的描述项目应比较齐全，信息资源的各个特征皆可以作为检索入口进行检索操作。

1. *元数据*

“元数据”一词是随着互联网的发展而出现的，目的是解决网络资源无序化的问题。现在一般把元数据这个词定义为关于数据的数据，或关于数据的结构化数据。本书认为，元数据是对信息资源进行描述、结构化并对之进行管理的工具。通过这种描述、结构化和管理，使得人和计算机能够识别、处理和利用各种信息资源。

元数据的基本功能是描述、检索和选择。在信息资源分类组织中，元数据用来描述一个信息对象的主题内容和位置，并建立各信息对象之间的关系，以实现信息资源的分类检索。用户

在进行信息检索时，首先要到信息管理部门的查询系统（分类检索系统）的元数据库中进行查找。通过元数据，用户可以了解该部门有哪些信息资源，判定信息资源的内容是否满足自己的需要，并确定自己需要的信息资源所在的位置，然后调度子系统根据元数据提供的位置获取所需信息。

长期以来，由于元数据本身较强的实践性，其定义很繁杂，学术界对此一直争论不休。较具代表性的定义如下。

（1）元数据是关于数据的结构化的数据。这个概念突出了元数据的结构化特征，从而使元数据作为信息组织的方式与全文索引有所区分。

（2）元数据是与对象相关的数据，此数据使其潜在的用户不必预先具备对这些对象的存在或特征的完整认识，它支持各种操作，用户既可能是程序，也可能是人。

（3）元数据是对信息包的编码描述，元数据的目的在于提供一个中间级别的描述，使得人们据此就可以做出选择，确定孰为其想要浏览或检索的信息包，而无须检索大量不相关的全文文本。

（4）元数据就是关于数据的数据，是指任何用于辅助进行数字资源的识别、描述和定位的数据。这意味着元数据能够采用多种形式，以不同的级别存在，为了满足不同的目的，可以通过多种方式生成。

按照元数据格式制定的时间及描述的信息资源类型，可将其分为传统元数据与现代元数据。其中，传统元数据的描述对象主要是各种传统的文献资源，表现形式为各种类型的书目、文献、索引等。有代表性的元数据格式为MARC（机读目录格式）等。现代元数据主要指为了对大量互联网信息资源进行描述、检索而产生的各种元数据，代表性的元数据格式有都柏林核心元数据元素集等。

2. 都柏林核心元数据元素集

1995年3月，由英国国家超级计算应用中心（NCSA）主持，52位来自图书馆界、计算机网络界的专家共同研究产生了都柏林核心元素集（Dublin Core Element Set），旨在通过建立一套描述网络资源的元素集合，来支持网络检索。从1995年到2008年9月，已经召开过16届会议，对该元数据元素集的结构和功能进行了不断补充和完善。

都柏林核心标准包括简单和限定两个层次。简单的都柏林核心包括15个元素，这15个元素都是可选的、可重复的，其定义如表6.1所示。限定的都柏林核心包括3个额外的元素，以及一组元素限定词，用于在资源发现时限定元素的语义要素。

由于信息环境向网络化、数字化的方向发展，对元数据的组织工具有了新的要求，包括：第一，可以在数字化、网络化的环境下应用；第二，适合网络信息资源的特点，能对它们实施有效的组织；第三，能被计算机自动处理和应用。要实现这三点，就要有一个共同的基础：采用机器可理解、可交换的方式来描述和表示信息。实现这一目标的方法就是采用机器标记语言。按照结构化、语义化的程度排列，网络标记语言有XML、RDF和HTML等。

表6.1　都柏林核心元数据元素集

元素名	中文名称	定义	备注
Title	资源名	赋予资源的名称	资源对象正式公开的名称
Creator	创建者	创建资源内容的主要责任者	创建者的实例包括个人、组织或某项服务系统。一般而言，用创建者的名称来标识这一条目

（续表）

元素名	中文名称	定义	备注
Subject	主题和关键词	有关资源内容的主题描述	如果要描述特定资源的某一主题，一般采用关键词、关键词短语或分类号
Description	说明	对资源内容的说明	说明元素可以包括但不限于以下部分：摘要、目录、以图形揭示内容的资源
Publisher	出版者	使资源成为可获得状态的责任者	出版者的实例包括个人、组织或某项服务系统。一般而言，用出版者的名称来标识这一条目
Contributor	其他责任者	对资源内容做出贡献的其他责任者	其他责任者的实例包括个人、组织或某项服务系统。一般用其他责任者的名称来标识这一条目
Date	日期	与资源本身生命周期中的一个事件相关的日期	一般而言，日期应与资源的创建或可获得的日期相关。建议采用的日期格式应符合ISO 8601[W3CDTF]规范，并使用YYYY-MM-DD的格式
Type	资源类型	有关资源内容的特征和类型	资源类型包括描述资源内容的一般范畴、功能、流派或聚类层次的术语。建议采用来自于受控词表中的值（如都柏林核心的资源类型指导性草案[DCT1]），为描述的物理或数字表现形式，应使用格式（FORMAT）元素
Format	格式	资源的物理或数字表现形式	格式可以包括资源的媒体类型或大小，如因特网媒体类型表[MIME]定义了计算机媒体格式
Identifier	资源标识符	在特定范围内给予资源的一个明确的标识	对资源的标识采用符合某一正式标识体系要求的字符串或数字。例如，统一资源标识符（URI），资源定位符（URL）等都是正式的标识体系
Source	来源	对一个资源的引用，当前资源源自这一引用资源	
Language	语种	描述资源知识内容所使用的语种	
Relation	关联	对相关资源的参照	
Coverage	覆盖范围	资源内容所涉及的范围	典型的范围包括空间位置，时间段、管辖范围
Right	权限	资源本身所拥有或被赋予的权限信息	

3. 资源描述框架

资源描述框架（Resource Description Framework，RDF）是由W3C于1999年开始开发的一种数据模型，用于表示网络资源的元数据，目前已形成多个W3C推荐标准。

RDF数据模型为元数据的定义和使用提供了一个抽象的概念化框架，基本的RDF数据模型由如下三类对象组成。

（1）资源，由RDF描述的任何事物都可以称为资源，如一本书、一个作者、一个网页、多个网页集合等。每个资源都有一个统一资源标识符（URI）。

（2）属性，也称性质，用于描述某一资源的特定方面、特征、属性和关系。每个属性都有其特定的含义，定义其允许值、可描述的资源类型及与其他属性之间的关系。

（3）陈述，也称声明，由资源、资源已定义的属性和该属性的值构成。陈述的这三个独立部分分别称为主体、谓词和客体。其中客体可以是另一个资源或文字。

下面以一个实例来说明RDF模型。

例 6.1　Diane Hillmann是资源http://dublincore.org/documents/2005/11/07/ usageguide/的创建者，则该语句的三个部分如下。

（1）主体：描述资源，其URI是

http：//dublincore.org/documents/2005/11/07/usageguide/

（2）谓词：描述资源的属性，在此是“创建者”。

（3）客体：表示属性值，在此是指“创建者”的值，即“Diane Hillmann”。

也可以使用以资源为节点的有向图方式来表示，其中资源和属性值都是以节点表示的，属性用有向弧表示（见图6.1）。

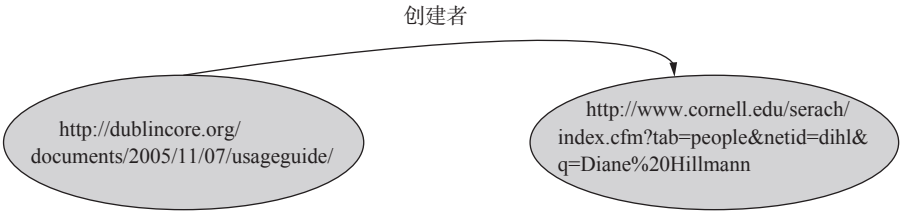


图6.1　例6.1的资源有向图表示

若客体本身也是一个资源（或称结构化实体），则客体也是以椭圆节点来表示的。若该客体是匿名的，则该节点为空节点，若该客体有URI，则在节点中给出URI。

例6.2　Diane Hillmann（E-mail为dihl@cornell.edu）是资源http://dublincore.org/documents/2005/11/07/usageguide/的创建者，如图6.2所示。

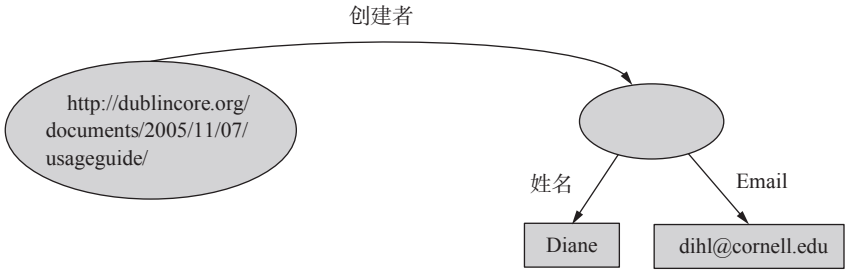


图6.2　例6.2的资源有向图表示

例6.3　在例6.2中，增加Diane Hillmann的信息所在：

http://www.cornell.edu/search/index.cfm?tab=people&netid=dihl&q=Diane%20Hillmann，则可以将图6.2改为图6.3所示的形式。

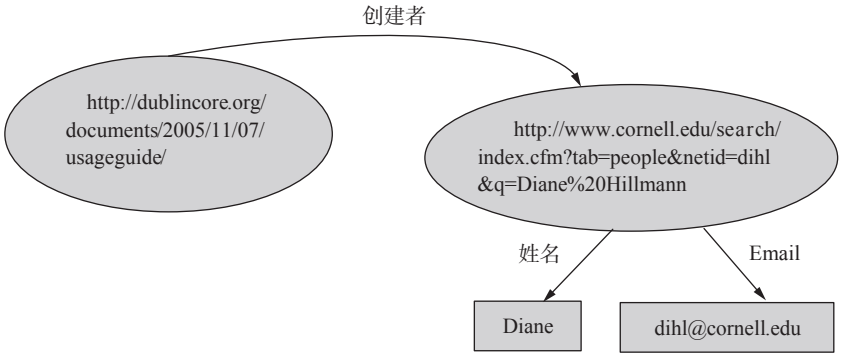


图6.3　例6.3的资源有向图表示

4. XML

网络的蓬勃发展，使得数字化文献在网络中快速传播和应用，不但常规的组织机构能够利用网络发布和收集文献，而且个人也能够利用网络发布自己的文献。然而，针对网络文献同样存在组织和有序化问题，在某种程度上，这一问题还更为突出和亟须解决。实现网络文献的有效组织，需要从文献的内容、结构和显示三个方面着手。为了能够恰当、准确地表示这三个方面的内容，人们提出了标记语言的方法。标记语言的基本思想是将文献需要加注的部分置于不同类型和名称的标记中，这些标记不仅能够标注文献的内容和结构，也能标注其表现形式。XML是一种重要而广泛使用的标记语言。

XML的语法规则既简单，又严格。这些规则很容易学习，也很容易使用，但只要文档中稍有违反XML规则的地方，XML解析器就会报错。图6.4是一组学生数据的XML文档。

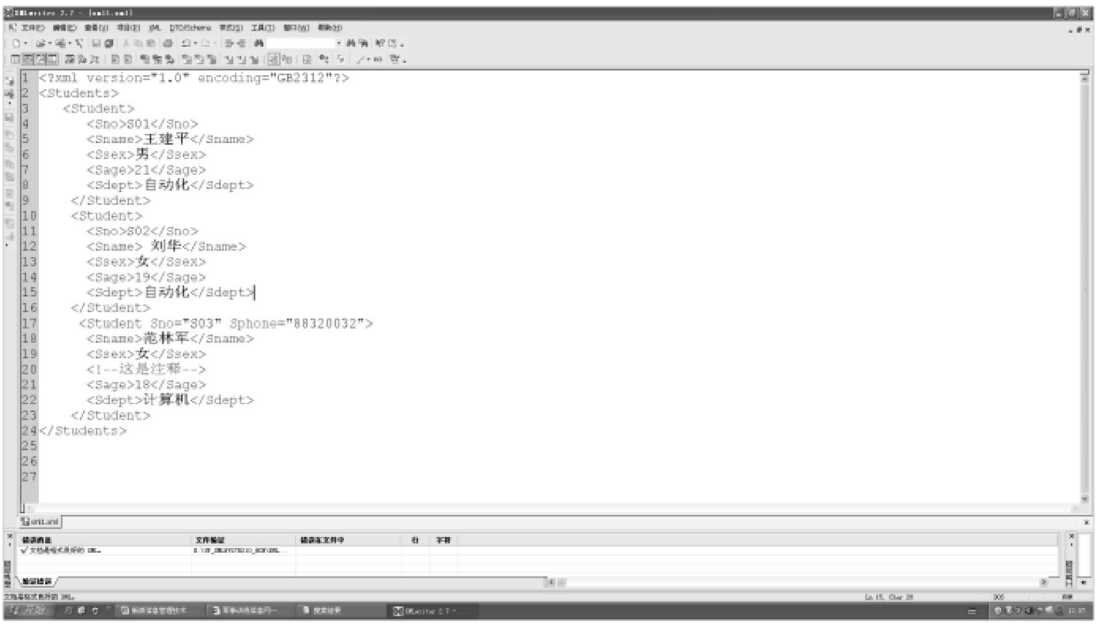


图6.4 学生数据的XML文档

XML文档分为两个部分：序言和文档元素。序言既可以只包含一个XML声明和一个注释，也可以包含其他的元素，如文档类型定义（DTD）。

XML文档是以元素为基本单位进行描述和管理的，即以<标签>开始，以</标签>结束。每组成对的标记及其标记的文档就称为XML中的元素。一个格式良好的XML文档需要具备以下条件。

（1）必须以XML声明开头

XML文档都必须以XML声明开始，声明中可以使用version、encoding、standalone三个属性。其中version属性用于指明XML文档所使用的版本，如XML1.0。每个XML文档都必须具有version属性，否则就是无效的。encoding属性用于指明XML文档编码使用的字符集。如果没有指定该属性，则XML处理器以unicode编码格式进行分析。目前存在多种语言编码集合，如简体中文编码GB2312、繁体中文Big5。然而，为了让XML文档内容能够处理多语种而不出现乱码问题，一般将encoding属性设置为UTF-8或UTF-16。standalone属性用于指明XML文档有没有使用外部标记声明，属性值为“yes”或“no”。如果standalone属性值为“yes”，则表明这是一个独

立的XML文档，它不需要其他文档就可以单独使用，在其内部没有任何引用外部资源的命令；standalone属性值为“no”，表明这不是一个独立的XML文档，在其内部引用了其他文档或资源，无法独立使用。

(2) 有且只有一个根元素

良好格式的XML文档中有且只有一个根元素。文档中的所有其他元素都是根元素的子元素。根元素的开始标签要放在所有元素之前，根元素的结束标签要放在所有元素之后。

(3) 开始标记与结束标记必须成对出现

XML中的标记有开始就一定有结束，如<Student>是开始标记，</Student>是结束标记，它们两个必须成对出现，否则XML解析器将会报错。

(4) 区分大小写

与C++、Java等语言一样，XML严格区分大小写字母，如<Student>和<student>是两个不同的标记，<Student>和</student>也不是一个成对的标记。书写时应该养成良好的习惯，例如全部使用大写，或者全部使用小写，或者首字母大写。

(5) 所有元素必须正确嵌套

XML标记之间不得交叉，所有标记必须有规则地按次序出现和结束。若要XML解析器不报错，就必须按照XML规范书写，例如：

<Student><Sno>S01</Sno></Student>

(6) 特殊字符的替代符

在XML中，一些类似“<”的字符拥有特殊的意义。如果把字符“<”作为XML元素中的内容就会发生错误，因为XML解析器会把它当成一个新元素的开始。例如，为了表达Sage小于20而用如下的元素表示，XML解析器就会报错。

<example>Sage<20</example>

为了避免这个错误，就需要用一个替代符“<”来代替“<”字符

<example>Sage< 20</example>

在XML中，有5个特殊字符的替代符号，如表5.2所示。

表5.2　XML中的特殊字符

替代符号	特殊字符	特殊字符解释
<	<	小于
>	>	大于
&	&	and
'	'	单引号
"	"	双引号

6.2.2　信息标引

信息标引是对信息资源内容特征进行主题分析并对主题分析结果赋予检索标识的过程，其目的是通过标引把信息资源与用户联系起来，使用户能在浩瀚的信息资源中全面、准确、迅速、便利地查找到特定的信息。信息标引是建立信息资源检索系统和进行信息检索的关键环节。

6.2.2.1 信息分类标引

1. 分类标引准备工作

(1) 选择分类法。分类法的优劣直接影响到类分信息资源的质量。因此，选择一部好的分类法至关重要。2005年以来，随着《军分法》作为国家军用标准的颁布实施，该分类法以其完整的军事科学体系设置和强大的信息分类组织与检索功能在军队信息管理部门开始得到应用。因此,《军分法》是目前类分军事信息资源比较适用的一部分类法。

(2) 熟悉分类法。选择好分类法后，就要进一步熟悉并学会使用这部分类法。一般情况下，可以从纵、横两个方面了解该分类法的结构体系。

纵：了解该分类法的编制原则是根据什么基本理论组织起来的，包含多少大类，大类的次序，以及每个大类的结构等。

横：了解分类表中类目的含义，弄清它包括什么，不包括什么；该类目与其他类目的关系等。如要选择《军分法》进行分类标引，就要对《军分法》的编制规律有所了解。如类目划分时采用了多维揭示的方式，设置了交叉类目、交替类目、类目参照，以及分类主题一体化的组织方式。

(3) 制定使用本。分类法的使用本，是指具体单位根据自己所藏的信息资源情况和用户需要，在允许的范围内，对通用分类法进行适当的调整、补充、说明后，确定下来作为分类标引最后依据的本子。

对于一个具体的信息管理部门来说，对《军分法》的使用要求也不尽相同。为此，有必要在本单位分类细则的基础上，制定适合本单位性质任务和所藏信息资源规模以便于用户使用的本子。

2. 分类标引工作流程

(1) 分类查重

分类查重是分类标引工作的第一步。要查明待标引的信息资源是否已被本单位、本系统或其他单位和系统标引过，有无标引成果可以直接采用或作为参考。

(2) 特征分析

特征分析是分类标引工作的第二步。对没有现成标引成果可以采用的待编信息资源，需予以标引，为此就要进行特征分析。特征分析又称主题分析，是指对标引对象进行内容特征分析，有时也包括某些必要的外表特征分析。进行特征分析一般有以下几种方法：

- 分析题名（书名、刊名或篇名）
- 阅读内容提要、浏览目录等
- 浏览正文
- 借助参考工具或请教专家

(3) 主题提炼及选择

主题提炼及选择是分类标引工作的第三步。在分析查明了信息资源的中心内容及主要特征后，还要深入分析反映信息资源主要特征的各组成要素，将其归纳成若干主题概念，再结合本单位的需求选择哪些主题因素予以标引，哪些不予以标引。

(4) 辨类与归类

辨类与归类是分类标引工作的第四步。在确定了作为分类标引对象的信息资源的主题概念后，即可根据信息资源主要主题的学科属性及分类标引规则，并结合本单位的需求，在分类法中选定与之相符的类目。

(5) 给号

给号是分类标引工作的第五步。所谓给号就是对辨类、归类结果具体给出分类号码的过程。在给号当中，特别注意的是类目号码的配置问题。当某种主题归入某一类目后，往往还要进行不同层次的细分。给号时，应将复分号置于主类号之后，同时还要注意各种复分号与主类号组配时的先后顺序，以保证最后给出的分类标识的准确性。

(6) 标引记录

标引记录是分类标引的第六步。标引记录包括两方面内容：一是标引成果著录，即将标引所得的检索标识按规定格式予以记载；二是每标引完一条新的信息资源记录，都要对标引中遇到的重要问题及处理结果在规范文档中做出相应的记录。

(7) 审校

审校是分类标引的第七步。为了保证分类标引工作的质量，在分类标引工作完成后，必须设专人进行审校。

3. 分类标引案例

依据前面阐述的分类标引的步骤，下面将介绍几种不同类型的信息资源分类标引过程与分类结果。

印刷型军事信息资源分类标引实例

如果将《中国战争动员百科全书》进行分类标引，则具体步骤如下。

① 分类查重。经过查重工作，发现此书并没有被某单位标引过，因此可以对其进行分类标引操作。

② 特征分析。通过对该书内容的分析，可知本书为军事专科百科全书，外表特征表现为参考工具书的特征。

③ 主题提炼及选择。本书的主题可提炼为：中国、战争、动员。

④ 辨类与归类。依据前面的三项步骤，可将《中国战争动员百科全书》首先归类于《军分法》的第Z类——军事信息综合资源，再进一步细分，可属于Z11类——军事百科全书类，再依据《军分法》进一步分类，可分为Z1188类——军事专科百科全书，再依据世界地区表组配复分，标引为Z1188(2) 。

⑤ 给号。根据上一步的分类结果，可将提炼出的主题“战争动员”的分类号附加于本类分类号之后，用组配方法组合，可得出《中国战争动员百科全书》的标引号为：Z1188(2) :H。

⑥ 标引记录。

⑦ 审校。

军事网络信息资源分类标引实例

如果将《军网网站资源搜索》进行分类标引，则具体步骤如下。

① 分类查重。经过分类查重工作，发现此书并没有被某单位标引过，因此可以对其进行分类标引操作。

② 特征分析。经过特征分析，可知该类资源为运行在军事训练信息网上的搜索引擎，作用是向用户提供军训网信息资源链接以及检索途径的检索工具。

③ 主题提炼及选择。本书的主题可提炼为：军事、目录、网络。

④ 辨类与归类。基于前三步的工作，可将该类资源归为Z大类——军事信息综合资源中的Z31——军事目、索引类，再进一步细分可归为Z3131类。

⑤ 给号。通过上一步的归类工作，再结合该资源的网络特性，因此可加入综合复分表中特种载体的复分号08，可最终得出《军网网站资源搜索》的标引号为：Z3131.08。

6.2.2.2 信息主题标引

1. 主题标引的概念

主题标引是依据一定的主题词表或主题标引规则，赋予信息资源语词标识的过程。具体而言，主题标引是在主题分析的基础上，以一定的主题词表或主题标引规则作为依据，将信息资源中具有检索意义的特征转化成相应的主题词，并将其组织成表达信息资源内容特征的标识的过程。按照是否使用主题词表，主题标引可分为两类：受控标引和自由标引。

2. 主题标引方式

主题标引方式是根指据信息资源的特点和使用需求而确定的标引和揭示信息资源主题的形式。主题标引方式一般分为以下几种类型。

(1) 整体标引。整体标引（亦称浅标引）是一种概括信息资源基本主题内容的标引。整体标引的对象可以是图书、论文、标准、档案或其他信息资源类型。这种标引只揭示信息资源中具有检索价值的整体性主题，不揭示其涉及的各种从属性主题内容。例如，对《信息组织概论》一书进行整体标引时，只需对“信息组织”这一主题内容进行标引即可。如果某一信息资源同时涉及两个或两个以上整体主题，则应根据情况，按照多主题的标引要求进行标引。例如，对于《分类法与主题法》这一文献，因其涉及两个整体主题，可标引为如，分类法+主题法。

(2) 全面标引。全面标引（亦称深标引）是一种充分揭示信息资源所论及的所有有检索价值的主题的概念标引。这种标引不仅要求揭示信息资源论述的整体主题，而且要求揭示符合检索系统要求的所有主题概念。

例如,《文献自动分类因果推理技术的实现》一文，结合《中图法》和《中国分类主题词表》，分析其内容和结构，形成字段分类因果推理规则，实现文献的自动分类，从而解决了目前自动分类的瓶颈问题。

(3) 对口标引。对口标引（重点标引）是一种只揭示信息资源中适合本专业需要的主题内容的标引。例如，“泰国的电子行业和汽车行业”这一主题，电子行业的研究单位可以用“电子行业”、“泰国”两个主题词进行对口标引；汽车生产研究单位可以用“汽车行业”、“泰国”两个主题词进行对口标引。

(4) 综合标引。综合标引是一种以丛书、多卷书、论文集、会议录、标准汇编、档案的案卷等为单位进行的概括性标引。综合标引除揭示信息资源的主题内容以外，一般还应根据情况对众多资源的类型进行必要的揭示。例如,《计算机科学丛书》在以整套书为单位进行标引时，除对整体内容“计算机科学”标引外，通常还对“丛书”这一类型进行必要的揭示，应标引为：计算机科学+丛书。

3. 主题标引的步骤

(1) 查找利用已有的标引成果。也就是说，要查明被标引的信息资源是否已被标引过，有无标引成果可以直接采用或作为参考。

(2) 主题分析。如果没有现成的标引成果可利用，则要进行标引。为此，需要对信息资源进行主题分析。基于人工的主题标引，要在充分了解信息资源内容及研究对象的基础上，对主

题的类型、结构及其构成要素等进行深入分析，对有检索意义的主题概念进行概括、提炼和选择。基于自动标引的主题分析则是，在信息资源中抽取表达主题的自然语词的方法，如词频统计分析、语词位置加权等。

(3) 主题概念的转换。用标引语言（标题语言、叙词语言等）的标识（标题词、叙词等）表达主题概念的过程称为主题概念的转换。在人工标引中，主题概念转换首先要辨识标引工具（标题表、叙词表）中的相应标识（标题词、叙词）的含义，然后选择恰当的标识。对于自动标引中的赋词标引，其主题概念转换是由计算机将文献中能表达主题的词与主题词进行相符性比较而完成的，将自然语言转换为主题语言。

4. 主题标引记录

主题标引记录是指在规定的载体上，按照一定的格式，对主题标引的结果和标引中所处理的一些重要问题做出记录。一般而言，在卡片或书本式索引中，主题标引的结果记录在卡片或书本式目录排检位置。在计算机检索系统中，则记录在文档的相应字段。

5. 审核

主题标引结果的审核是主题标引的最后一道工序，也是保证主题标引质量的一个重要环节。审核包括以下几方面的内容：信息资源主题的提炼是否全面、准确，是否遗漏隐含的主题和有潜在用途的主题；标引方式是否符合检索系统和信息资源类型的要求；选用的主题词是否符合标引规则等。

6.2.3　信息整序法

6.2.3.1　分类法

1. 分类的含义与特点

信息资源分类，是指根据信息资源的内容属性和其他特征，将其分门别类地、系统地组织和揭示的方法。一般来说，信息资源分类是以知识分类或学科分类为基础，结合信息资源各种载体的实际编制的类目体系，它与知识分类既有相同点，又有所不同。

为了有效地对信息资源进行分类，需要了解信息资源的本质属性与非本质属性。信息资源的本质属性是指载体上所记录的科学知识内容，它从根本上体现信息资源的价值和使用价值；信息资源的非本质属性是指与本质属性相对应的属性，一般体现在信息资源的形式特征上，如载体、语种等。

信息资源的分类是一种从主题内容角度组织和揭示信息资源的方法，是分类方法在信息组织中的应用。

2. 分类法的类型

从编制方式的角度来划分，信息资源分类法通常可分为三种类型：等级列举式、分面组配式和列举–组配式。

(1) 等级列举式分类法

等级列举式分类法是一种传统的分类法类型，是将所有的类目组织成一个等级系统，并且采用尽量列举的方式编制的分类法。这种分类法通常将类目体系组织成一个树形结构，按照划分的层次，逐级列出详尽的专指类目，并在以线性形式显示时，以缩格表示类目的等级关系。

等级列举式分类法是目前国内外使用得最普遍的分类法形式，比较著名的等级列举式分类法有：国外的《杜威十进分类法》(简称《杜威法》或DDC)、《美国国会图书馆图书分类法》(简称《国会法》或LCC）等；我国的《中国图书馆分类法》(简称《中图法》）等。

(2) 分面组配式分类法

分面组配式分类法是依据概念的分析与综合原理，将概括信息资源内容与事物的主题概念组成“分面–亚面–类目”的结构体系，通过各分面内类目之间的组配来表达信息资源主题的一种信息资源分类法，也称为组配分类法或分析–综合分类法。与等级列举式分类法相比，分面组配式分类法放弃了详细列举类目体系的做法，采用以简单概念组成复合类目的方式。分面组配式分类法的核心是分面。分面又称组面，简称面，所谓面就是按某种分类标准（分类特征）产生出来的一组类目。

分面组配式分类法的最典型代表是印度著名图书馆学家阮冈纳赞所创制的《冒号分类法》。

(3) 列举–组配式分类法

列举–组配式分类法又称为半分面分类法，是在等级列举式的详尽类表的基础上广泛采用各种组配方式的分类法。

列举–组配式分类法兼有前面两种分类法的特点。其优点是：以列举式类表为基础，具有一定的直观性，同时广泛采用组配方法，基本上可以达到与分面组配式类表同等的标引水平。其不足是：在相关类目的修订或改进方面需要投入大量的资源和精力，在实现类目之间的组配时，需要使用分面组配式分类法的多种辅助符号或号码进行标记，最终导致分类标引方面的标记程序比较复杂，标记符号也显得冗长。

3. 分类法的结构体系

分类法的结构体系一般由类目体系、标记符号、说明与注释、类目索引四部分组成。

(1) 类目体系

类目体系是根据类目内在关系和一定原则建立起来的类目集合，是分类法的主体，也是分类语言进行词汇控制的主要依据。它一般是以知识分类为基础，按照信息资源分类的实际需要而建立的。类目体系由主表和复分表构成，主表一般在基本部类的基础上，由基本大类、简表和详表等构成。

复分表就是将主表中按相同标准划分某些类所产生的一系列相同子目抽取出来，配以特定号码，单独编制成表，供主表有关类目进一步复分用的类目表。复分表也称附表、辅助表、副表、共性区分表，是分类法的重要组成部分。

(2) 标记符号

标记符号也称号码，是分类体系内类目的代码或代号，通常称为分类号。标记符号的作用在于：一方面表示类目在分类体系中的位置；另一方面表示类目的排列顺序。

(3) 说明与注释

说明与注释是分类法中帮助用户了解和使用分类体系的重要组成部分，一般用来说明类表的编制原则、类目体系的特点及使用方法等。说明与注释通常包括编制说明、大类说明和类目注释等三种形式。

(4) 类目索引

类目索引也称分类表索引，是从类目名称以字顺为途径查找相应分类号的工具，为分类法的有效使用提供了便利。

类目索引主要有两个作用：一是通过将分类体系的系统排列转变成字顺排列，可以从表达主题的词语出发找到相应的分类号，克服了类目查找的困难；二是便于用户查找分类表中被分散在各个学科门类的有关同一事物的类目及分类表中未列出的有关新概念。

类目索引主要包括三种类型：直接索引、相关索引和主题词索引。直接索引是一种直接通过类名或同义词查找对应类目的索引。相关索引是一种不仅可以从主题名称出发查找对应的类目，而且可以将被分类体系分散的该主题各方面的类目加以集中的工具。

6.2.3.2　主题法

1. 主题法的含义

“主题”一词，在不同的语境中有不同的解释。在信息组织方面，主要指信息资源所论述的主要对象，包括事物、问题、现象等。那些经过选择的用来表达信息资源主题的语词称为“主题词”。“主题法”是指直接以表达主题内容的语词作为检索标识，以字顺为主要检索途径，以参照系统等方法揭示词间关系的标引和检索信息资源的方法。

主题法实际上包含两个含义：第一，是指信息资源的主题整序方法，即用语词标识标引信息资源和组织检索系统的方法；第二，是指主题语言。也就是说，主题法包含主题标引和主题语言。主题标引是指对信息进行主题分析，用主题语言表达分析出的主题，赋予信息资源主题标识的过程；而主题语言是一种检索语言，标题词、元词、叙词等主题词就是主题语言的主体。

2. 主题法的特点

(1) 直接以语词作为主题标识

主题法不像分类法那样，以一种抽象的号码系统作为分类标识，而是直接选用自然语言中的语词作为主题标识。例如，“土壤生态学”这一主题，在《中图法》中的分类标识为S154.1；但在主题法中，可直接用“土壤生态学”作为主题标识，比分类标识直观。

(2) 以字顺排列作为主要检索途径

虽然主题法往往也采用范畴（分类）、词族（等级）等方式对主题词进行组织，但字顺方式始终是它的主要排检依据。我国的主题检索系统通常是根据汉语特点、按照拼音或笔画笔顺进行排检的。因此，在使用主题法检索时，只要知道检索对象的名称，就可以按相应的字顺帧排检方式进行查找，方便快捷。在采用机检系统的情况下，一般可以直接输入语词，由计算机进行查找，不必像使用分类法那样必须预先了解主题所属的学科，因此通用性较好。

(3) 以主题为中心集中信息资源

分类法由于受学科体系的限制，从不同学科角度研究同一对象的信息资源是被分散在各知识门类的，主题法则没有这一限制，而是直接从主题对象的角度揭示信息资源，这一特性是由主题标识和字顺排列决定的。以论述葡萄的文献为例，在分类法中，关于葡萄的栽植、葡萄的酿制、葡萄的贸易等主题，一般应按学科分别归入农业科学、工业技术、经济等不同学科门类。而在主题法中，通过语词标识和字顺排列，这些有关葡萄的各种信息资源可以直接在“葡萄”这一主题标识集中予以揭示。

(4) 通过参照系统等方式揭示主题词之间的关系

为了在采用字顺排列的同时有效地揭示主题概念之间的关系，主题法发展了完备的参照系统，通过在主题词下设置用、代、属、分、参等多种参照项，建立起“隐蔽的分类体系”。此外，一些还具有词族索引、范畴索引、轮排索引等辅助索引，从各种不同的角度体现主题之间

的关系。通过上述各种形式的结合，在主题词之间建立起充分的语义联系。当然，各种主题系统中对词间关系的揭示状况是不平衡的，就整体而言，其主题之间关系的揭示不如分类法。

6.3 信息检索

信息检索即查找所需信息的过程。或者说，信息检索就是用户根据需求，借助于检索工具，从信息集合中找出所需信息的过程。广义的信息检索包括信息的存储和检索两个过程，是指将信息按照一定的方式组织和存储起来，既能根据用户的需要找出其中的相关信息，同时又包括了“存”和“取”两个基本环节，全称为“信息存储与检索”。而狭义的信息检索仅指根据用户的需要找出其中的相关信息的过程，也就是我们平常所说的信息检索。

6.3.1 信息检索语言

检索语言又称信息语言、标引语言、信息存储与检索语言等，是为沟通文献标引与文献检索而编制的人工语言，也是连接信息存储和检索两个过程中标引人员与检索人员双方思路的渠道，是用于文献标引和检索提问的约定语言。编制检索语言不但能够保证不同的标引人员描述文献特征的一致性，而且能够保证检索提问标识与文献标引词的一致性。这就是说，概念与所要描述的事物一一对应，尽量减少一词多义或多词一义的现象，要使其在该检索系统中具有单义性。

按照不同的划分标准，可将信息检索语言分为不同的类型。这里主要从构成原理的角度来讨论检索语言的类型。依照此标准，可将信息检索语言划分为表述文献信息外部特征的语言和表述文献信息内部特征的语言。表述文献信息外部特征的语言有题名语言、著者语言、号码语言等；表述文献信息内部特征的语言有分类检索语言（分类法）、主题检索语言（主题法）。

6.3.2 信息检索工具

6.3.2.1 信息检索工具简介

检索工具是用来报道、存储、查找信息的工具，是汇集各种信息并按特定的方法加以描述和编排形成的。因此，检索工具是一种特定类型的出版物，与普通文献的主要区别在于它是一种工具书刊，是专供人们查找特定信息的，虽然也具备可读性，但不是供人们系统地阅读的。从检索工作的实际需要看，一切用于查检特定信息的信息源都是检索工具。检索工具的特点表现为以下几方面。

1. 编排特殊，便于查检

检索工具是按某种体例编制的，是专供查考特定信息，而不是供系统地阅读的书刊。所以说，可检性是检索工具的最明显特征，有利于人们查考。

2. 信息密度大

检索工具作为人们解释疑难的工具，要求内容准确，易于查检，要求在有限的载体内容纳尽可能大的信息量，供人们进行查找。

3. 概括性强

检索工具的编排要求在信息的选择上，采用概括的手法，从大量的信息源中筛选、比较、浓缩、提炼，选择具有代表性的信息加以收录；在信息的描述上，采用概括的手段，既能表达信息，又简明扼要。

随着网络技术与计算机水平的不断发展，传统的以纸质工具为检索对象的手工检索系统，因为其效率低、实效差等缺陷已逐渐被淘汰，取而代之的是高效的计算机检索系统。计算机检索系统以数据库为检索对象，检索效率高，已成为用户获取信息的主要手段之一。

根据存储内容不同，可将计算机检索系统分为文献信息检索系统、事实信息检索系统、数值信息检索系统、图像信息检索系统和多媒体信息检索系统等五类。

根据工作方式不同，可将计算机检索系统分为单机检索系统、联机检索系统、光盘检索系统和网络检索系统等，其中网络检索系统因具有检索灵活、及时性好且检索费用低等优势，已成当今用户检索网络信息资源的主要手段。对网络信息资源进行检索的方式主要有浏览式检索、直接访问检索、搜索引擎等三种方式，搜索引擎是最常用的一种方法。

6.3.2.2　搜索引擎概述

搜索引擎指的是一种在网络上应用的软件系统，它以一定的策略在网络上对大量信息进行搜索和发现，然后对这些信息进行提取、排序、质量分析并提供信息的存储、检索等服务。从用户的角度来看，搜索引擎提供了一个网页页面的方式，用户通过填写搜索条件（词或短语）提交搜索请求，然后搜索引擎将返回一个与用户搜索条件相关的信息列表（列表中的每一个条目都代表了一个网页链接）。

搜索引擎需要对用户的搜索请求即时响应，因此搜索引擎不可能在用户提交搜索请求时才去网络搜索资源，而是预先保存了大量网页，以数据库的形式保存在系统中，此时的搜索只是在系统内部进行而已。另外，搜索引擎无法判断用户的背景，因此给出的搜索列表并不一定满足用户的需求。所以，现在大多数搜索引擎都提供了个性化搜索服务，即根据用户历史的搜索行为对结果进行排序，并将用户可能最关心的那些信息放在列表的前列。

搜索引擎是一个“网络导航工具”，与用于提供图书馆馆藏信息的目录系统相似，搜索引擎本身并不提供任何实际的Web文档，而仅提供关于网页的信息。搜索引擎为所采集的每一个网页建立一条记录，记录包括对网页的简单描述、标题以及实际网页所在服务器的URL等信息，这些记录的集合就构成了索引数据库。搜索引擎通过对索引数据库的采集与调用来实现网络导航功能。

6.3.2.3　搜索引擎的类型

随着搜索引擎数量的急剧增加，其种类也越来越多。目前，主流的搜索引擎按内容组织方式主要分为以下几类。

1. 目录型搜索引擎

目录型搜索引擎以人工或半自动方式将Internet网站按类别编排形成一份目录，各类别之下排列着属于这一类别的网站的站名和网址链接，有些搜索引擎通过人工方式还提供了各网站的信息摘要。这类搜索引擎由于加入了人工维护，信息较为准确，导航质量较高，用户只需遵循分类体系即可准确方便地找到所需的信息。用户通过浏览层次型的目录来寻找相关的信息资源，目录按一定的主题分类体系组织，并辅之年代、地区等分类。用户一般采取逐层浏览目

录、逐步细化来寻找合适的类别直至具体资源，可以不依靠关键词进行查询。但其缺点是需要人工介入、信息维护量大、受用户主观影响以及信息分类的交叉导致许多内容的重复等。目前，国内典型的目录型搜索引擎有搜狐、新浪等。

2. 全文检索型搜索引擎

全文检索型搜索引擎通过使用大规模数据库来收集、组织和存储互联网资源，搜索引擎将用户输入的关键词在数据库中进行匹配和关联，然后将匹配的结果以列表清单的形式返回给用户。这种方式更新较快且方便直接，可以使用逻辑关系组合关键词，通过添加各种语法规则比如数据类型、范围、时间等来精确定位，因而可准确检索满足特定条件的网络资源。虽然这种方式返回的信息量较大，但是用户很难找到所需的内容，检索结果缺乏准确性，包含的可用信息少。另外语法规则过于烦琐，一般用户也很难做到熟练掌握。国外具有代表性的全文搜索引擎有Google、Yahoo等，国内有百度、中搜、搜狗等。它们都是通过从Internet上提取各网站信息（以文字为主）而建立的数据库，检索与用户查询条件匹配的相关结果集，然后按一定排序方式将结果集返回给用户。

实际搜索时，我们可以在文字框中输入要查找的字、词或短语，再点击按钮，搜索引擎便会查找相关的站名、网址和内容提要，返回查到的内容。按钮的名称，一般是“开始”，也有的起名为“查找”、“搜索”、“查询”等，我们所输入的那个字、词称为关键词，因此有时也称之为关键词检索。

6.3.2.4 搜索引擎的工作原理

搜索引擎实际上就是一个网络应用软件系统。基本原理就是通过网络蜘蛛或网络机器人定期地在互联网上查找，发现新的网页，将它们取回放到本地数据库中，用户的查询请求可以通过查询本地数据库获得。搜索引擎大致由四个功能模块组成：网页收集器（信息提取）、索引器、查询（检索）器和用户接口，这四个模块构成了搜索引擎工作的四个阶段，其基本结构如图6.5所示。

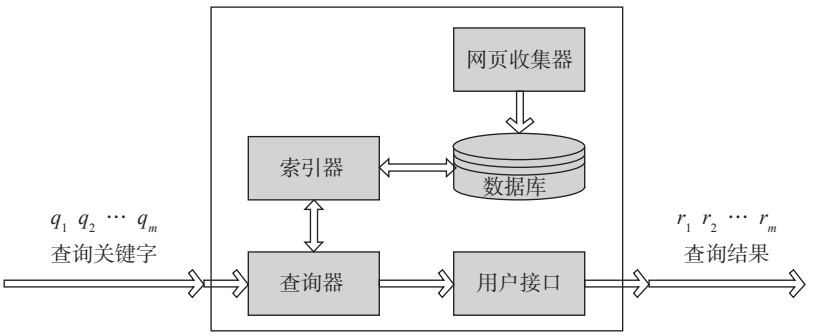


图6.5 搜索引擎的基本结构

1. 网页收集器

用户利用搜索引擎检索信息实际上是在搜索引擎的数据库中进行检索，而不是实时的查询。因此，如何保证搜索引擎数据库的信息量和新鲜度，是网页收集器要考虑的主要问题。网页收集器需要在浩瀚的互联网中漫游，发现和搜集网页资源。它通常是一个计算机应用程序（比如网络蜘蛛或网络机器人），它会尽可能多、尽可能快地搜集各种新页面，由于互联网上的

信息更新很快，因此还需定期更新已搜集过的旧信息。每个搜索引擎都会定期收集互联网上的网页，并更新原有页面，比如Google在一段时间是每隔28天更新一次。这种方式实现简单，但是新鲜度不高。同时，网络中页面的生命期也是有限的，据统计50%页面的平均生命周期为50天。因此可以使用增量更新方式，即每次只收集新出现和上次收集后更新的页面，并将自上次收集后已经无效的链接去除，这样可以极大地降低数据库的数据量，提高检索性能。

2. 索引器

索引器的优劣直接影响搜索引擎的质量。索引器从网页收集器获取的页面中抽取信息，从中抽取索引项用于表示页面，以及生成页面数据库的索引表。这一步相当复杂。据北大天网提供的搜索结果显示，2005年底我国的网页总数约为10.8亿个，去掉重复后网页总数约为3.7亿个，相当于网络中每一个页面，平均会有另2个页面与其相同，这是因为网络中存在大量转载和复制。这会极大增加索引器的工作量，重复的网页只需做一个索引项，这需要判断页面的相似度。根据天网统计，网页文档源文件的大小通常约为其内容大小的4倍。另外，许多网页内容上较为随意，网页中有很多与内容无关的信息，比如广告、导航条、网站说明等。如何从源文件中提取内容，这给信息查询带来了很大的挑战。在这一步，自然语言处理扮演了重要角色。以中文为例，需要先去除无关的HTML标记，然后进行中文分词（将中文序列分成若干个词，词是汉语的基本表达单元），同义词处理，去除“停用词”（是指那些没有实际意义的词，比如“的”、“是”、“在”等）。经过上述步骤后，整个页面可以用若干个词的序列来表示。

索引器还有一个很重要的作用就是体现网页的重要程度。由于搜索引擎数据库中与搜索关键词相匹配的结果集较大，如何将用户最想要的结果列在返回结果集的前端，这就涉及每个搜索引擎的核心排序算法。在判断这个网页与另一个网页相比哪个更重要时，通常采用科技文献的做法，即“引用越多越重要”。比如Google就是根据此思想，如果一个网页URL被更多网页引用，则表示这个网页比较重要，体现在排序中就会列在返回结果集的前列。

3. 查询器

查询器根据用户的查询在索引库中快速检索，进行相关度匹配，对检出的结果进行排序，返回相应的网页列表给用户。这包括：首先需要对用户查询进行理解，将其转换为服务器检索使用的信息，比如进行同义词和语义转换的处理；然后，根据用户查询，从索引库中检索出结果集；再次，进行相关度判断，根据特定的排序算法，对输出结果进行排序。其中，相关度判断通常采用向量空间模型，用户查询与页面文档表示为向量，相似度则体现为用户查询与页面文档向量间的夹角余弦；最后，将相关度大于阈值的页面集按照相关度逆序返回给用户。当然，搜索引擎的排序并不一定与用户的需求完全符合。

4. 用户接口

用户接口模块为用户提供可视化的查询输入和结果输出界面，方便用户输入查询关键词、显示查询结果、提供用户相关性反馈机制等，目的是方便用户使用搜索引擎，多方式地检索出有效的信息。同时，用户接口的设计和实现必须适应人类的思维习惯，比如用户接口提供了各种查询运算符（and、or、+或−）、出现位置（如标题、内容或网站）、域名范围（如.edu、.org或.cn）等限制查询条件。在查询输出界面中，将查询器检出的结果集形成一个线性文档列表，其中包含页面的标题、简介和链接地址等信息，由页面生成系统将搜索结果的链接地址和页面内容、摘要等内容组织起来返回给用户。

6.3.3 信息检索技术

掌握和正确使用检索技术有助于提高检索效果的提升。针对应用领域的不同，主要可分为一般检索技术和网络信息资源检索技术。

6.3.3.1 一般检索技术

1. 常规法

常规法又称为工具法，是指通过各种信息检索工具查找信息的方法，可分为：顺查法、倒查法和抽查法等。

2. 引文法

引文法又称追溯法。引文法是利用原始文献后面所附的参考文献进行追溯查找文献的方法。该方法简单易行，但工作量大，检索出来的文献有时比较陈旧，或者有些文献与检索课题相距甚远，所以该方法一般是在检索工具不齐全或者无检索工具时使用。

3. 循环法

循环法又称综合法。该方法实际是对前两种方法的综合运用，即在信息检索时，先利用检索工具查出一批相关信息，然后通过筛选，选择出与课题关系特别密切的文献信息，按照其后所附的参考文献进行追溯查找，分期、分段地交替进行，从而获得大量的相关文献信息。

6.3.3.2 网络信息资源检索技术

1. 布尔逻辑检索

布尔逻辑检索是依据布尔逻辑算符来完成的，而规定检索词之间的逻辑关系的算符则称为布尔逻辑算符。布尔逻辑算符包括逻辑“或（OR）”、“与（AND）”和“非（NOT）”。大多数检索工具具有布尔逻辑检索功能，有的允许进行“逻辑与”、“逻辑或”、“逻辑非”三种逻辑运算，有的只能进行两种或一种逻辑运算。

2. 截词检索

利用检索词的词干或者不完整的词形进行检索，是指把检索词截断，取其中的一部分片段检索，用截词符“？”或“*”代替词干以外的字符。“？”和“*”称为通配符，其中，“？”代表一个任意字符，“*”代表任意个任意字符。

3. 位置检索

位置检索是用一些特定的算符（位置算符）来表达检索词与检索词之间的关系，并且可以不依赖叙词表而直接使用自由词的检索方法。位置算符又称邻接算符，适用于两个检索词以指定间隔距离或指定顺序出现的情况，例如以词组形式表达的概念、彼此相邻的两个或两个以上的词、被禁用词或以特殊符号分隔的词等。位置算符是调整检索策略的一种重要手段，按照两个检索词出现的顺序和距离，可以有多种位置算符，而且对同一种位置算符，检索系统不同时规定的位置算符也不同。位置算符主要有以下几种。

（1）(W)——With

(W)表示在此算符两侧的检索词必须按此前后顺序排列，顺序不允许颠倒，而且两个检索词之间不允许有其他的词或字母，但可以有空格和标点符号。例如，Information(W)retrieval可检索出information retrieval和information-retrieval。

（2）(*n*W)——*n*Word

(*n*W)表示在此算符两侧的检索词之间允许插入*n*个（最大数量）实词或虚词（非用词），两个检索词的词序不能颠倒。例如electronic(1W)resources，可检索出electronic resources，electronic information resources。

（3）(N)——Near

由(N)连接的检索项在记录中出现的顺序可以调换，即查找两个连在一起的单词。例如junior(N)high可检索出junior high，high junior。

（4）(*n*N)——*n*Near

(*n*N)表示两个词位置可以颠倒，两个词之间插入词的最多数目是*n*个。例如，information(3N)retrieval，可检索出information retrieval，retrieval information，retrieval of information，retrieval of law information，retrieval of Chinese law information等，information和retrieval两个词之间最多可插入3个词。

（5）(F)——field

(F)表示在此运算符两侧的检索词必须同时出现在文献记录的同一字段内，如出现在篇名字段、文摘字段、叙词字段、自由词字段，但两个词的词序不限，且两个词之间可间隔若干个词。

（6）(S)——Sentence

(S)表示在此运算符两侧的检索词只要出现在文献记录的同一个子字段内（例如在文摘中的一个句子就是一个子字段），此文献即被命中，两个词的词序不限，且两个词之间可间隔若干个词。

4. 字段检索

字段检索是指将检索词限定在数据库记录中的一个或几个字段范围内进行查找的一种检索方法。在检索系统中，数据库设置的可供检索的字段通常有两种：表述文献信息内部特征的基本字段、表述文献信息外部特征的辅助字段。基本字段又称主题字段，包括篇名、文摘、叙词、自由标引词等字段。辅助字段又称非主题字段，包括文献记录的作者、文献类型、语种、出版年等字段。

6.4　信息导航技术

古往今来，人类从甲地移动到乙地时，往往会借助导航工具的帮助。例如道路旁边的路标、海上的灯塔、指南针、GPS等，这些均属于物理空间的导航。同理，在网络空间的移动也离不开信息导航。

6.4.1　基本概念

一些学者认为信息导航是指借助一定的信息引导用户，使之能够快速、正确地到达其预定目标的行为。在物理世界的导航中发挥作用的认知地图、空间认知等在超文本空间中同样能发挥作用，只是形式略有不同。与物理空间相比，网络信息空间具有如下特点：超链接构成的超维空间；动态变化的多用户空间；异化的信息结构空间。网络信息导航是一个整合了“位移”和“路径发现”的行为，指通过相关时间、空间、路径、情景等信息的辅助，促进用户在网络信息空间中进行位移的过程，位移的目标是实现一定的路径发现任务。

6.4.2 技术实现

1. 分类导航

分类导航是指从分类的角度，将相同主题的信息集中在一起，形成知识体系，向用户展示不同主题之间的包含和从属关系，因而可以有效地支持用户在某一学科领域或专业范围进行研究。常见的分类导航主要有如下三种。

(1) 直接采用现成的分类法体系进行设计。目前常用的分类导航主要基于“中图法”的类目对信息资源进行组织。事实上，许多学科领域已形成关于本学科的分类体系。这种体系相比于综合的分类体系而言，更能融合该学科的特点。在特定的情况下，如在专业知识仓库的分类导航中，应该使用专业分类体系对相关内容进行组织。

(2) 自行开发一套更符合系统特点的分类体系。系统的用户群和学科领域不同，所提供的资源和服务也不尽相同。系统可以针对本系统用户的学习和研究需求，开发出更符合用户期望的知识导航方法。

(3) 综合方法。部分学术导航系统同时采用了上述两种方法实现分类导航。

2. 知识关联链接

由于文献与文献之间总是存在着一定的联系，不同的文献之间的各自参照和引用，形成了一个庞大的文献关联网络。从一份文献出发，去了解与该文献有关的其他文献信息，是用户进行研究的一种非常重要的方式。通过对文献中的知识内容进行浓缩、提炼，编成摘要等，使用户可以花费较少的时间和精力，获得较多的信息，为用户提供多种知识关联链接，以揭示文献之间的相互关系，进而帮助用户构建关于某一领域的知识地图。

3. 检索结果聚类

检索结果聚类是指从改变排序方式的角度对检索功能进行改进。通过文本聚类技术对检索结果自动分组，聚集成相关文献的子集合，并通过文献特征的比较和提取，形成新的分类结构。用户可根据类目，选择更为相关的检索资源，因此既能帮助用户迅速剔除自己不需要的文献，又能帮助用户发现单纯使用排序输出检索结果时很难发现的有用文献。现有的检索系统通常可以按相关度和时间对检索结果进行排序，这种单纯依靠某一指标排序的方式，容易导致其他与指标耦合度较低但符合用户真实需求的文献排在结果集中比较靠后的位置，以至于用户未能浏览到相关的文献。

4. 目录指南

目录指南又称为主题树方式，组织信息资源的方法是将信息资源按照某种事先确定的主题，分门别类地加以组织，用户通过浏览的方式层层遍历，直到找到所需信息的线索，再链接到相应的网页。这种方式主要是通过主题及主题之间的关系来组织数据库中的资源。用户利用主题导航，可以了解到关于某一主题或事物的所有信息，还可以通过该主题发现学科和专业之间的相互联系，促进科学研究和创新。其优点为信息的专题性较强，信息质量高，且能较好地满足族性检索的要求。用户按照规定的范围和分类体系，逐级查看，按图索骥，目的性强，查准率高。目录指南方式屏蔽了网络资源系统相对于用户的复杂性，提供了一个基于树浏览的简单易用的网络信息检索与利用界面，并且具有严密的系统性和良好的可扩充性，在建立专业性或示范性的网络信息资源体系时，显示出了其结构清晰、使用方便的优点。目录指南方式通常

需要完备的后控词表，因而对于专指性强且有学科主题词表的专业数据库而言，主题导航的建立显得更为容易。许多专业的数据库中都采用了主题词表来实现检索的后控。

5. 基于资源外部特征的导航

这种方式主要是利用信息资源的外部特征实现，如图书文献资源，可利用出版机构导航、首字母导航、参考文献导航等。我们常用的期刊论文库、学位论文库、会议论文库等多种文献特征数据库大多属于此类。

6.4.3　典型应用

1. 国外应用

20世纪90年代初，国外已经开始建立相关学科或主题的信息导航，如美国加州大学伯克利分校参考馆员1990年制作的LII（Library Internet Index），就是将图书馆组织信息资源的方法应用于学科信息门户的典型。90年代中期以后，以主题网关（Subject Gateway）形式出现的学科信息导航系统建设不断涌现。在学科资源主题导航应用中，国外学科主题门户站点对资源的分类通常采用三种方式：通用分类法、单学科分类法和自建分类法。同时，这些网站还使用主题词表，提供相关的主题词以提高检索的效率。如：SOSIG使用了国际十进分类法（Universal Decimal Classification） 和HASSET（Humanities and Social Science Electronic Thesaurus） 叙词表标引；Business and Education商业与教育门户网站使用杜威十进分类法（Dewey Decimal Classification，DDC）的精简版，是一种使用单学科分类法的门户站点；BUBLLink采用了DDC分类法和LCSH（Library of Congress Subject Threading）；英国爱丁堡大学自然科学与工程学院虚拟图书馆EEVL和EELS使用EI（Engineering Information）分类法和叙词表。

IPL（Internet Public Library）主要集成和整合网站资源，其资源的选择过滤、分析评价、导航体系构建等工作都由人工完成，信息集成的质量高，但资源覆盖范围有限，存在资源链接更新不及时等问题。

随着数字图书馆的发展，信息资源的集成范围不再仅仅局限于网络资源，也开始包括本地馆藏资源。例如，阿拉莫斯实验室将本地馆藏资源、授权商业资源、网络开放获取资源等统一按学科进行集成和整合，同一学科内再按数据库、电子期刊、网站、书目信息等分类，支持基于学科的浏览服务。当前，阿拉莫斯实验室这种信息集成揭示模式已被众多图书馆采纳。

2. 国内应用

目前，我国比较大的信息导航系统有（CNKI中国国家知识基础设施）中国期刊网、万方数据资源中心网站、重庆维普资讯公司中外文科技期刊数据库网站，以及百度、中文雅虎、搜狐等公共资源导航系统。

同方公司作为CNKI的建设单位，已成为国内著名的数据资源加工商，其开展数据资源服务的特点为如下。

（1）对数据资源，提供了期刊导航、基金导航、作者单位导航、内容分类导航、博士学位授予单位导航、硕士学位授予单位导航、会议主办单位导航、会议论文集导航、报纸导航、出版社导航等。从这些导航方式来看，主要是按照数据资源类型种类以及资源的外部特征进行的导航，经过这些导航方式过滤的数据基本都是依照关键词匹配的模式，搜索结果依然很庞大，无法满足用户的需求。

（2）CNKI中国期刊网导航服务，针对所收录的期刊，分别提供了专辑分类、期刊主办单位、期刊所在地区、期刊首字母、期刊排名、期刊发行频率等导航方式，供用户浏览。用户既可以直接在期刊导航中查询某一种刊物，也可以选择其中一种方式查看相应分类下的期刊。CNKI则在单库检索和跨库检索中分别使用了自定义的专辑分类导航和中图分类法导航。

（3）CNKI的医学资源库使用了《医学主题词表》（中文2002年版）和《中国中医学主题词表》，实现了主题词–关键词–分类号之间的关联控制，建立了医学资源的主题导航系统。

万方数据公司在万方知识服务平台上，依据文献类型、学科、日期、刊名等，（对检索结果）提供了聚类。该平台在期刊导航时采用自定义的分类体系，而在学科导航时则使用“中图法”的类目划分。目前，该平台的导航系统主要提供如下三种知识关联链接方式。

（1）检索节点链接，即在一篇文献中往往导航了多个可用检索的检索节点，如作者、单位、刊名、年期等，这些信息通常出现在检索结果的题录或摘要中。用户在浏览当前文献时，能够获得同一作者的其他著作，全面了解该作者的研究方向和参与课题的相关进展。检索节点的链接对应于该平台提供的HotLink功能。

（2）基于引证关系的资源链接，即文献之间的关系主要通过引证关系来体现，通过引文去追踪学科之间的种种内在联系，可以找到一系列内容相关的文献，进而了解某一学科的研究动态、发展情况，以及该学科的核心作者群等。基于引证关系的资源链接包括：引文链接、引证文献链接、共引文献链接、同文献被引链接。

（3）相似文献链接，即相似文献之间在内容上非常相关。为弥补基于引证关系的资源链接不足，在知识关联链接设计中，还将相关的信息资源提供给用户，利用动态聚类算法，通过分析不同文献之间的内容和主题，将那些在内容上与当前文献最接近的文献聚集起来，提供关于这些资源的链接。

6.5 信息推荐技术

无论广义上的信息管理，还是狭义上的信息管理，它的基本宗旨都是为了更好、更高效地发挥信息资源的价值，充分利用好信息资源。然而，信息服务的进一步发展面临着两个重大挑战。首先是如何提高计算机深入理解信息意义，准确捕获用户需求的能力，这个方面的突破将大幅提高信息服务的质量，提升用户的使用感受。其次是如何利用信息、服务以及用户之间的交互提取语义信息，建立起信息和用户的语义联系。个性化服务的主要工作集中在日志挖掘和用户需求模型构建方面，信息推荐主要是一种分析结果的应用，不同系统可能采用不同的用户建模和推荐技术，同一个系统也可以采用多种用户建模和推荐技术。随着互联网的不断发展，信息量急剧膨胀，面对内容多样化、形式多元化的网络信息资源，如何利用现代化信息技术手段提供有效的信息服务，使用户能将零散的隐性知识快速转化为具有使用价值的系统化的显性知识，成了现代信息服务迫切需要解决的问题。因此，本节主要以网络个性化信息服务为切入点具体介绍相关服务内容与所用技术。

6.5.1 概念与特点

从狭义上讲，信息推荐技术就是把信息自动地送到用户面前的技术，即实现“信息找用户”。举例来讲，电视台每天播放电视节目，观众打开电视就可收看电视节目，并可随时选台，这就是推荐；手机用户每天以短信方式收到的天气预报、股票信息等定时发送，也是推荐。

1. 信息推荐技术的概念

信息推荐技术是指依靠网络公司，通过一定的技术标准或协议，从网上的信息源或信息制作商那里获取信息，通过固定的频道向用户发送信息的新型信息传播系统。它能够根据用户对信息的需求，有针对性和目的性地将用户所需信息主动送达用户。

信息推荐技术的基本工作流程是：用户填写订阅单，其中包括用户个人档案、所需信息类型以及需要推荐的时间等相关内容，并提交给信息提供商；然后信息提供商按用户的订阅单收集相关信息，并由推荐服务器推荐给用户。客户终端获取信息完毕后告知用户可读取信息。关于用户获取内容的渠道，既可以是直接将信息源中的信息本身送给用户，也可以是只将有关信息的目录或索引通知发送给用户，由用户根据通知去查询相应的信息。

2. 信息推荐技术的特点

信息推荐技术改变了互联网上信息访问的方式，将用户搜寻信息变为有目的地接收信息。这不仅改变了信息流动的方向，而且通过信息流量的减少降低了因特网的负载。其特点如下：

(1) 灵活的用户设置；
(2) 内容定制文件；
(3) 无缝连接；
(4) 持久文件传输；
(5) 有效利用带宽；
(6) 新旧内容自然衔接；
(7) 通知方式灵活；
(8) 安全性好；
(9) IP协议组对有用的信息进行分类和管理，以固定的信息频道进行播发，能极大地提高信息开发利用的程度和管理的力度。

当然，目前推荐技术也存在一些问题，如不能确保发送成功、无信息状态跟踪、针对性差、信源任务重等。但是，随着推荐技术的不断发展和完善，并与“信息拉取”技术取长补短，未来能够形成“智能信息推拉”技术，具有广阔的应用前景。

6.5.2　服务形式

信息推荐服务的形式可分为两种：基于Internet的信息推荐和基于智能数据库的信息推荐。

1. 基于Internet的信息推荐

基于Internet的信息推荐可以采取以下几种形式。

(1) 通知。推荐技术最基本的形式是一个简单的通知，如电子邮件等。针对这种形式，读者可控制它通知的形式、时间间隔等。通知并不具备很强的交互性和强制性，对资源和信息流量的要求不高。

(2) 提要。通过提要可实现查看Web页或其他信息源，寻找需要匹配的信息、并向读者传递信息。提要有很多后台进行的处理活动，不仅是给读者每天一次的报道，它的活动还要受查找条件的制约，这些后台的处理活动与后台的联系是不可预测的。

(3) 自动拉出。这种形式有一组可供读者经常查看的Web页。自动拉出将获得所有这些Web页，并保存起来供读者以后阅读。自动拉出可以获得许多资料，读者还可以通过电子邮件接收这些资料，或至少知道这些Web页是为自己编制的。

(4) 自动推荐。自动推荐能根据自身的刷新时间发布信息。读者预定推荐信息服务，需要在Web页上连续收听广播。在一般情况下，这种服务要求在读者终端上安装特殊的客户机软件，定期发出更新请求。这种推荐交互性强，读者可以选择需要查看的信息流，系统也可以精选发送给读者的信息，或者试探读者感兴趣的其他信息。

(5) 频道式推荐。这是目前普遍采用的一种模式，它将某些页面定义为浏览器中的频道。读者可以像选择电视频道那样接收感兴趣的信息。

(6) 网页式推荐。在一个特定网页内将所推荐的信息提供给读者，如某企业、某组织或某个人的网页。

(7) 专用式推荐。采用专门的信息发送和接收软件，信源将信息推荐给专门读者，如机密的点对点通信。

2. 基于智能数据库系统的信息推荐

基于智能数据库系统的信息推荐主要可以采取以下几种形式。

(1) 操作式推荐（客户式推荐）。由客户进行数据操作，启动信息推荐。当某客户对数据进行操作时，把修改后的新数据存入数据库后，即启动信息推荐过程，将新数据推荐给其他客户。

(2) 触发式推荐（服务器推荐）。由数据库中的触发器启动信息推荐。当数据发生变化后，出现增加、删除、修改操作时，触发器启动信息推荐过程。

6.5.3 用户建模

个性化系统的核心是用户建模，即对用户信息需求的模型表示。用户建模是个性化系统的核心。概括地说，用户建模是对用户信息的一种表示，主要包括三个问题：对什么建模、如何建模、如何维护模型。

用户兴趣建模是一个复杂的过程，既包括用户相关信息的收集与挖掘，也包括用户兴趣的提取和表示，以及长期兴趣、近期兴趣、即时兴趣的组织与维护。以上各个方面相互联系、相互作用，共同构成一个完整的用户建模体系。例如，用户兴趣模型采用的表示方法决定了信息收集时必须记录的信息，同时也决定了长期兴趣、近期兴趣、即时兴趣的组织结构，因此必须统一在同一个框架下才能深入地探讨用户建模过程。对什么建模，与系统需求密切相关，根据不同的应用目的有不同的建模对象，包括知识、兴趣、背景、任务、特性和上下文，例如网络教学系统更注重对学生“知识”水平的建模，信息推荐系统更注重对用户兴趣的建模，而面向不同类型设备的服务系统则更注重对设备的特性进行建模。为了论述的简单性，接下来假设用户模型反映的仅仅是用户的“背景知识”和“兴趣要点”，并称之为“用户兴趣模型”。

建模方法包括三个方面：(1) 如何获得用户的有用信息；(2) 用户模型的表示方法问题；(3) 如何将前两个方面结合起来，从而产生出用户的模型。

建模是一个过程，包括很多方面，不同应用背景下的建模过程也会不同，但一般至少包括三个方面：(1) 用户行为的模型和表示，并根据该模型记录用户的具体访问行为，产生出用户行为日志；(2) 根据用户的行为模式评价用户对所访问的信息项的关注程度；(3) 根据用户所访问的信息内容对用户的兴趣进行提取和量化评价，构建用户的兴趣模型。这三个问题密切相关，前一个问题均为解决后一个问题的前提或基础。

由于用户信息会发生转变，因此需要提供一个动态跟踪机制来捕获这些变化，从而能分辨出不同兴趣之间的差异。用户兴趣模型的变化主要表现在兴趣度的衰减和强化两方面，并由此

表现出用户兴趣的迁移。从整体用户行为历史来看，每个兴趣点被访问的次数是单调增长的，因此相应的兴趣度都在被强化；而衰减的目的是表达一个兴趣点如果在最近一段时间内没有被访问，则其兴趣度应该被降低；而“最近一段时间”又是一个很模糊的概念。

为了正确捕获用户的兴趣模型，可将用户兴趣分为如下三种情况。

（1）长期兴趣。反映了用户稳定的长期信息需求，这些兴趣会随着时间推移积累成广泛的兴趣点，这将导致信息推荐发散在多个主题上，从而缺乏针对性。

（2）近期兴趣。反映了用户最近一个时间段内的信息需求趋势，其基本目标是根据用户的近期访问行为，在用户长期兴趣中选择几个作为用户近期的关注焦点，从而克服基于长期兴趣进行推荐的问题。

（3）即时兴趣。反映的是用户在与系统交互过程中的实时信息需求，既可能是某个稳定兴趣的体现，也可能是与长短期兴趣均无关的临时信息需求。

尽管三种用户兴趣互不相同，但它们之间又有着内在的联系。概括地说，即时兴趣是近期兴趣的累积基础，而近期兴趣是长期兴趣的累积基础；即时兴趣和近期兴趣能够反映用户需求的动态变化，而长期兴趣则体现了用户较为稳定的信息需求。

鉴于三种兴趣之间的密切联系，应该综合考虑一个完整的用户模型框架，其必须能够反映三者之间的递增关系，同时也能突出强调用户兴趣的动态与静态特征。

用户模型的动态特征捕获值得进一步深入研究。目前主要采用的方法是通过近期兴趣来建立用户当前的兴趣点，从而捕获用户兴趣的变化。

一般而言，近期兴趣和长期兴趣在建模过程上是相似的，只是前者将时间限制在某个特定阶段内。因此，用户模型的动态特征捕获方法尽管有一定的效果，但其模型基础并不完善，表现在以下三个方面。

（1）长期兴趣与近期兴趣并没有统一在一个框架下；

（2）长期兴趣与近期兴趣之间的关系并不明确；

（3）并未涉及衰减和强化两方面的机制问题，因此也不能刻画用户兴趣的迁移模式。

在用户兴趣模型的表示上，主要包括基于关键字、基于语义网络、基于本体等三种方法，它们在所包含的语义信息多少方面是递增的，而用户兴趣度一般均以量化的数值来表示。显然，语义信息越多，用户模型越准确，应用就越灵活。例如，用户对拦截导弹和航空母舰感兴趣，则有如下几种方式。

（1）基于关键字方法的表示，则是（拦截导弹，10.5）且（航空母舰，12.3），其中的数值表示兴趣度。

（2）基于语义网络方法的表示，是（拦截导弹，10.5）且（航空母舰，12.3）且partof（拦截导弹，航空母舰），即在关键字表示的基础上增加了关键字之间的语义关系；partof（拦截导弹，航空母舰）的含义是“拦截导弹是航空母舰的组成部分”。

（3）基于本体方法的表示：拥有一个丰富的知识库，其中包括一些知识。例如：IsA（拦截导弹，巡航导弹），其含义是“拦截导弹是一种巡航导弹”。Instanceof（“爱国者导弹”，拦截导弹），其含义是“爱国者导弹是拦截导弹的一种特例”。根据这些知识可以推断出用户对巡航导弹和“爱国者导弹”都感兴趣。显然，语义越丰富，在信息推荐时可推荐的内容就越丰富、越具体。

6.6 云平台技术

信息服务系统是集成各种数据资源和软硬件环境的大规模分布式系统，其运行平台必然是一个综合网络、操作系统、数据库和各种运行软件的异构环境。云平台技术作为一项新兴的技术，无论是在信息存储还是信息服务领域都有着广阔的应用前景。

6.6.1 基本概念

大量计算机构成的资源池称为云。云是一些可以自我维护和管理的虚拟计算资源，是指提供资源的网络，通常是一些大型服务器集群，包括计算服务器、存储服务器和宽带资源等。其利用高速网络的传输能力，将数据的处理过程从个人计算机或服务器转移到网络上的大量计算机集群构成的资源池中，使各种应用系统能够根据需要获取计算能力、存储空间和各种软件服务。云计算是一种新型的超级计算方式，以数据为中心，是一种数据密集型的超级计算。云中的资源在使用者看来是可以无限扩展的，并且可以随时获取，按需使用，随时扩展，按使用付费。用户可以动态申请部分资源，支持各种应用程序的运转，能够更加专注于自己的业务，有利于提高效率、降低成本和技术创新。云计算技术是由分布式处理、并行处理和网格计算发展而来的，并融合了近年出现的虚拟化、Web 2.0等热点技术，是这些计算机科学概念的商业实现与应用创新。云计算作为一种新兴技术，已受到全球关注。

云计算的基本功能的实现取决于两个关键因素：数据存储能力和分布式计算能力。因此，云计算的云可以细分为存储云和计算云。存储云是一个大规模的分布式存储系统，对第三方用户开放存储接口，用户可以根据自己的需求购买相应的容量和带宽。计算云包括并行计算和资源虚拟化。

6.6.2 关键技术

1. 云计算关键技术主要包括数据存储、资源管理等技术

(1) 云计算环境下的数据存储技术

为保证高可用、高可靠和经济性，云计算采用分布式存储的方式来存储数据，采用冗余存储的方式来保证存储数据的可靠性，即为同一份数据存储多个副本。云计算系统需要同时满足大量用户的需求，并行地为大量用户提供服务。因此，云计算的数据存储技术必须具有高吞吐率和高传输率的特点。

(2) 云计算环境下的数据管理

云计算系统对大数据集进行处理和分析，向用户提供高效的服务，数据管理技术必须能够高效地管理大数据集。如何在规模巨大的数据中找到特定的数据，也是数据管理技术必须解决的问题。云计算对海量的数据存储、读取后进行大量的分析，数据的读操作频率远大于数据的更新频率，云中的数据管理是一种优化的数据管理。因此，云系统的数据管理往往采用数据库领域中列存储的数据管理模式，将表按列划分后存储。

(3) 提供简便、交互性好的编程接口模型

为了使用户能更轻松地享受云计算带来的服务，让用户能利用该编程模型编写简单的程序来实现特定的目的，云计算上的编程模型必须十分简单，必须保证后台复杂的并行执行和任务调度对用户和编程人员透明。

2. 架构分层

典型的云架构分为三个基本层次：基础设施层、平台层和应用层。这三种层次向上提供服务的方式包括公有云、私有云和混合云等三种类型。

（1）基础设施层

基础设施层是经过虚拟化后的硬件资源和相关管理功能的集合。云的硬件资源包括了计算、存储和网络等资源。基础设施层通过虚拟化技术对这些物理资源进行抽象，并且实现内部流程自动化和资源管理优化，从而向外部提供动态、灵活的基础设施层服务。

（2）平台层

平台层介于基础设施层和应用层之间，是具有通用性和可复用性的软件资源的集合，为云应用提供了开发、运行、管理和监控的环境。平台层是优化的云中间件，能够更好地满足云的应用在可伸缩性、可用性和安全性等方面的要求。

（3）应用层

应用层是云上应用软件的集合，这些应用构建在基础设施层提供的资源和平台层提供的环境之上，通过网络交付给用户。

3. 主要服务

云架构中的每一层都可以为用户提供服务，分别为基础设施即服务（Infrastructure as a Service，IaaS）、平台即服务（Platform as a Service，PaaS）和软件即服务（Software as a Service，SaaS）。

（1）IaaS

IaaS交付给用户的是基本的基础设施资源。基础设施向用户提供了虚拟化的计算资源、存储资源和网络资源，这些资源能够根据用户的需求进行动态分配。

服务提供商架设出规模巨大的数据库及存储中心，提供存储空间、网络设施、带宽等，将资源网络化、虚拟化，并以服务的形式提供。实现海量存储，保证其安全性、隐私性和可靠性，而且要高效、低价、节省能源。

（2）PaaS

PaaS交付给用户的是丰富的云中间件资源，这些资源包括应用容器、数据库和消息处理等。因此，PaaS并不面向普通的终端用户，而是面向软件开发人员，使其可以充分利用这些开放的资源来开发定制化的应用。

（3）SaaS

SaaS交付给用户的是定制化的软件，即软件提供商根据用户的需求，将软件或应用通过租用的形式提供给用户使用，该软件通过网络交付给用户，用户无须在本地安装该软件的副本。用户只需要通过浏览器就能获得任何种类的服务，包括手持设备的专用浏览器。云计算架构下，用户自主决定最符合其利益的资源部署方式，从而在云和端之间实现平衡。通过云+端、云端互动，最大程度地利用云的功能实现用户的最佳体验。

6.6.3　典型应用

Salesforce公司是云计算中SaaS厂商的先驱。该公司网络应用软件平台Force. com可作为其用户自身软件服务的基础。Force. com包括关系数据库、用户界面选项、企业逻辑以及一个名为Apex的集成开发环境。程序员可以在平台的Sandbox上对他们利用Apex开发出的应用软件进行

测试，然后在Salesforce的App Exchange目录上提交完成后的代码。用户还可以通过Force. com搭建一个综合网站，自己只需通过HTML、JavaScript、Flex和CSS设计用户界面。

IBM推出了“改变游戏规则”的蓝云计算平台，为客户带来即买即用的云计算平台。IBM在全球建立了近20家云计算中心，仅在中国就建立了无锡、北京和山东东营等三个云计算中心。IBM的蓝云计算平台为用户搭建了可通过Internet访问的分布式云计算体系。它整合了IBM自身的Tivoli、VMWare虚拟化软件、Hadoop开源分布式文件系统，由数据中心、管理软件、监控软件、应用服务器、数据库以及一些虚拟化的组件共同组成。

Sun的Network. com将来自独立软件厂商和开放源代码开发社区的软件编成目录，并在线提供高性能计算应用程序，把它们作为云计算服务提供给用户。用户可以在平台上构建或部署自己的应用程序；如果用户同意，还可以把自己的应用程序共享出来，作为服务提供给其他用户。Sun云服务的核心是虚拟数据中心。

亚马逊公司最先推出云计算，主要是基于虚拟化技术，为用户提供通过网络访问的存储、计算机处理、信息排队和数据库管理系统等接入式服务。亚马逊云服务主要由4块核心服务组成：简单存储服务（Simple Storage Service，S3）、弹性计算云（Elastic Compute Cloud，EC2）、简单排列服务（Simple Queuing Service）以及Simple DB。其中，S3为用户提供无限量的数据存储；EC2让用户自行选择虚拟软件开发环境配置，包括内存大小、运算单位和存储空间。EC2和S3为企业提供计算和存储服务，收费的服务项目包括存储空间、带宽、CPU资源以及月租费。

Google是目前云计算最大的实践者。Google以学术论文的形式公开了其云计算三大法宝：GFS、MapRecluce和Bigtable，并于2008年4月推出了应用服务托管商用平台——Google应用引擎（Google App Engine，GAE）。开发人员可在此之上编写应用程序，用户可以使用定制化的网络服务。Google搜索引擎建立在200多个站点、超过100万台服务器的支撑之上，而且这些设施的数量正在迅猛增长。Google的一系列成功应用平台，包括Google地球、地图、Gmail、Docs等也同样使用了这些基础设施。

微软公司把云计算定义为云+端、软件+服务。2008年10月，微软公司推出云计算平台Windows Azure。在Windows Azure操作系统上，目前运行着Live Services、NET Services、SQL Services、Share Point Services和Dynamics CRM Services五大服务，作为微软公司下一代网络服务的基础。

思考题

1. 简述元数据的功能。
2. 分类标引的特点和作用是什么？
3. 基于XML语言，采用RDF框架，试描述网络资源中的某一多媒体资源。
4. 什么是合法的XML文档？它要满足哪些原则？
5. 信息导航有哪些关键技术？
6. 信息推荐技术用户建模的基本方法是什么？
8. 什么是云平台？有哪些关键技术？
9. 什么是检索语言？
10. 针对信息推荐的用户建模包括哪几个方面？

第7章　信息安全技术

由于信息在产生、分发、存储、处理各个阶段都可能面临各种各样的威胁，信息安全自古就受到学者、军事家和政治家的重视；特别是随着信息技术的飞速发展，信息安全面临的形势也日益严峻，因此必须不断更新和完善信息安全技术体系，提高信息安全技术水平，加强信息安全防护能力，以更好地适应信息技术发展和应用的需求。

7.1　信息安全基本概念

7.1.1　信息安全定义

信息安全是指保护信息系统的硬件、软件及相关数据，使之不因为偶然或者恶意侵犯而遭受破坏、更改及泄露，保证信息系统能够连续、可靠地正常运行，信息服务不中断。信息安全基本属性一般可归纳为如下五个方面。

(1) 保密性。确保信息只被授权人访问，免受非授权的泄密攻击，保证信息不泄露给未经授权的人。

(2) 完整性。保证所获取的信息和原始信息的一致性，防止信息被非法篡改。

(3) 可用性。确保授权的用户在需要时可以访问信息。换句话说，保证合法用户对信息和资源的使用不会被不正当地拒绝。

(4) 可控性。对信息及信息系统的安全性进行监控。这一点可以确保某个实体的身份的真实性，也可以对非法活动进行监控。

(5) 不可否认性。又称为抗抵赖性，保证信息行为人不能否认其信息行为。这一点可以防止参与某次通信交换的一方事后否认该次交换曾经发生。

7.1.2　信息安全威胁

所谓信息安全威胁，是指某些因素（人、物、事件、方法等）对信息资源及系统的安全使用可能构成的危害。造成信息安全的威胁既有主观原因，又有客观原因。分析归纳信息安全存在的威胁来源，可以归结为如下五个方面。

(1) 自然灾害、意外事故。因水灾、火灾及地震等自然灾害或人为破坏，或者令牌、身份卡被盗等意外事故所带来的信息安全威胁。

(2) 人为错误。由于操作人员安全配置不当造成的安全漏洞，安全意识差，用户口令选择不慎等带来的信息安全威胁。

(3) 系统设计问题。由于操作系统、网络协议及应用软件自身缺陷带来的信息安全威胁。

(4) “黑客”行为。通过对计算机及网络系统进行攻击而造成的信息安全威胁。

(5) 信息战。利用暴力破解、电磁干扰等手段造成的信息安全威胁。

在计算机网络中，常见的信息安全威胁具体有以下几种。

(1) 窃听。监听窃取信息资源和敏感信息。例如，对通信线路搭线监听，对局域网使用嗅探软件抓取数据包再解析信息，或者利用通信设备产生的电磁泄露获取有用信息等。

(2) 重放。出于非法目的，将所截获的某次合法的通信数据进行复制，而重新发送。

(3) 假冒身份。通过欺骗通信系统（或用户）达到非法用户冒充成为合法用户，或者特权小的用户冒充成为特权大的用户的目的。

(4) 抵赖。这是一种来自用户的攻击，比如否认自己曾经发布过的某条消息、伪造一份对方来信等。

(5) 计算机病毒。一种在计算机系统运行过程中能够实现传染和侵害功能的程序。

(6) 特洛伊木马。看起来正常的程序和资料，内嵌了恶意的程序，用于窃取用户的信息或实施破坏行为。这种伪装过的程序或资料称为特洛伊木马（Trojan Horse）。

(7) 后门。在某个系统或某个部件中设置的“机关”，使得在特定的数据输入时，允许违反安全策略。

(8) 拒绝服务攻击。黑客攻击信息提供方，使得正常用户无法获取访问信息。

(9) 业务流分析。通过对系统进行长期监听，利用统计分析方法对诸如通信频度、通信的信息流向、通信总量的变化等参数进行研究，从中发现有价值的信息和规律。

(10) 旁路控制。攻击者避开正常访问流程，利用系统的安全缺陷或安全性上的脆弱之处获得非授权的权利或特权。

7.1.3 信息安全保障体系

“信息安全保障体系”为信息系统安全体系提供了一个完整的设计理念，同时较好地诠释了安全保障的内涵。信息安全保障体系通常采用PDRR模型（见图7.1）。

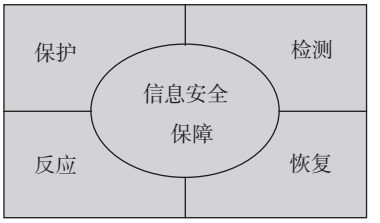


图7.1 信息安全保障体系

(1) 保护

所谓保护，就是指预先采取安全防范措施，迫使产生攻击的条件无法形成，让攻击者无法实施入侵信息系统的行为。保护属于被动防御，无法彻底地阻止各种针对信息系统的攻击行为。主要的安全保护技术包括网络安全技术、信息保密技术、操作系统安全技术、物理安全防护、访问控制技术及病毒预防技术等。

(2) 检测

所谓检测，是指根据有关的安全策略，采取相应的技术手段，针对可能被入侵者利用的信息系统的脆弱部分，进行具有一定实时性的检查，同时将结果形成检测报告。主要的检测技术包括入侵检测、恶意代码检测及脆弱性扫描等。

(3) 反应

所谓反应，是指对于破坏系统安全性的事件、行为、过程及时做出适当的相应处理，避免危害的后果进一步恶化，使信息系统蒙受的损失最小。主要的反应技术包括报警、跟踪、阻断及反击等相关技术。反击又可分为取证和打击，其中取证是根据相关法律法规搜集入侵者的犯罪证据，而打击是采用合法手段反制入侵者。

(4) 恢复

所谓恢复，是指当危害事件发生以后把信息系统恢复到原来的状态或比原来更安全的状态，降低危害对信息系统造成的损失。主要的恢复技术包括异常恢复、应急处理、漏洞修补、系统和数据备份及入侵容忍等。

7.1.4　信息安全系统设计原则

网络信息系统在设计过程中应该遵循以下原则。

(1) 木桶原则。设计安全系统要均衡协调所有安全模块的保护级别。假如，系统整体的安全性都比较高，但存在一处较弱的安全环节，也会影响整个系统的安全性。

(2) 整体原则。设计安全系统要有一整套安全防护、监测和应急恢复机制。

(3) 有效性与实用性原则。安全性的设计不能影响系统正常运行和合法用户的操作。

(4) 安全性评价原则。系统是否安全，没有绝对的评判标准和衡量指标，只能决定于系统的用户需求和具体的应用环境。

(5) 动态化原则。在设计安全的系统时要尽可能引入更多的可变因素以提高安全性，还要设计具有良好的扩展性的系统，允许用户自主增加新的安全模块。

(6) 等级性原则。必须根据实际需求，所设计的安全系统必须能达到一定安全层次和安全级别。这些安全标准由安全组织规定。国际上有一些统一的安全标准，下面将详细进行介绍。

7.1.5　安全标准

为了帮助计算机用户区分和解决计算机网络安全问题，评估信息是否安全，国际上和我国均制定了相应的标准。在信息安全系统评价方面，下面列出了一些重要的安全评估准则有：

(1) 可信计算机系统评估准则

可信计算机系统评估准则（Trusted Computer System Evaluation Criteria，TCSEC）由美国国防部和国家标准技术研究所制定，将信息的安全描述为4个方面，即安全政策、可说明性、安全保障和文档。标准综合这些方面对信息安全性进行评价，将不同的安全性能分为7个级别，从低到高依次为D、C1、C2、B1、B2、B3和A级。

(2) 信息技术安全评测准则

信息技术安全评测准则（Information Technology Security Evaluation Criteria，ITSEC）由欧洲共同体制定。与TCSEC不同的是，ITSEC把完整性、可用性与保密性作为同等重要的因素。ITSEC定义了从E0级（不满足品质）到E6级（形式化验证）的7个安全等级。

(3) 国际标准ISO/IEC 15408

国际标准化组织（International Organization for Standard，ISO）在1999年发布了IEC 15408“信息技术、安全技术、信息技术安全性评估标准”，也就是Common Criteria v2.1标准，目的是把已有的安全准则结合成一个统一的标准。该标准结合了TCSEC及ITSEC的主要特征，强调将安全的功能与保障分离，并将功能需求分为9类63族，将保障分为7类29族。它提供了安全服务与有关机制的一般描述，确定了在参考模型内部可以提供这些服务与机制的位置。

(4) 中华人民共和国国家标准系列

中华人民共和国国家标准系列简称国标，用GB表示。GB17895-1999《计算机信息系统安全保护等级划分准则》将信息系统安全分为5个等级，分别是自主保护级、系统审计保护级、安全

标记保护级、结构化保护级和访问验证保护级。主要的安全考核指标有身份验证、自主访问控制、数据完整性、审计、隐蔽信道分析、客体重用、强制访问控制、安全标记、可信路径和可信恢复等，这些指标涵盖了不同级别的安全要求。

7.2 密码技术

密码技术是信息安全的核心基础技术，能够提供机密性、完整性、真实性和不可否认性等属性。因此，密码技术在信息安全领域中占有重要地位，在信息安全中的应用将会不断拓宽。

7.2.1 基本概念

保密学是研究信息系统安全保密的科学，包括密码学和密码分析学两部分。密码学是对信息进行编码，实现隐蔽信息的一门学问，其主要目的是寻求保证信息的保密性和认证性的方法。密码分析学是研究分析破译的学问，其主要目的是研究加密信息的破译或信息的伪造。密码学和密码分析学相互独立、相互促进地向前发展。

(1) 加密与解密

数据加密过程就是通过加密系统把原始的数字信息（明文），按照加密算法变换成与明文完全不同的数字信息（密文）的过程。

解密过程为加密过程的逆过程，即将密文变换成明文的过程。

一个数据加密系统包括加密算法、明文、密文和密钥，密钥控制加密和解密过程。一个加密系统的全部安全性是基于密钥的，而不是基于算法的，所以加密系统的密钥管理是一个非常重要的问题。

在密码学里，约定用M表示明文，用C表示密文，用E()表示加密函数，用D()表示解密函数，用K表示密钥。加解密的过程可以用图7.2来表示。

明文：M
密文：C
加密函数：E　　加密：$E_k(M) = C$
解密函数：D　　解密：$D_k(C) = M$
密钥：K

图7.2　加解密过程示意

(2) 密码体制分类

密码体制从原理上可分为两大类，即对称密码体制和非对称密码体制。对称密码体制又称单钥、私钥或传统密码体制，非对称密码体制又称双钥或公钥密码体制。

对称密码体制是指信息的发送方和接收方共享一把密钥。在现代网络通信条件下，该体制的一个关键问题是如何将密钥安全可靠地分配给通信的对方，并进行密钥管理（见图7.3）。

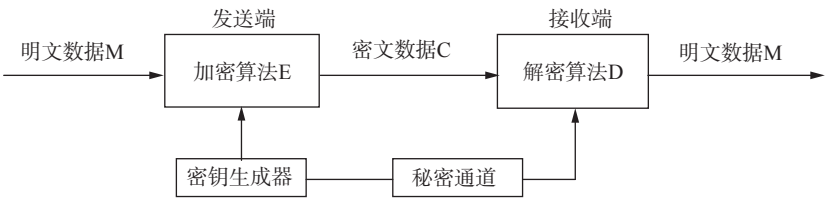


图7.3　对称密码体制

对称密码体制在实际应用中除了要设计出满足安全性要求的加密算法外，还必须解决好密钥的产生、分配、传输、存储和销毁等多方面问题。因为通信对象的多元性导致了一个用户必须拥有多个不同对象的密钥，方可安全可靠地进行通信，因此当网络中用户较多时，秘钥管理是一个复杂的问题。

非对称密码体制的最大特点是采用两个密钥将加密、解密分开。在双钥体制下，每个用户都拥有两把密钥，一个公开，一个自己专用。利用用户专用密钥加密，并利用该用户公开密钥解密时，即可实现一个加密的消息可被多个用户解读；利用用户公开密钥加密，并利用该用户专用密钥解密时，即可实现传输的信息只被一个用户解读。前者常用于数字签名，后者常用于保密通信（见图7.4）。

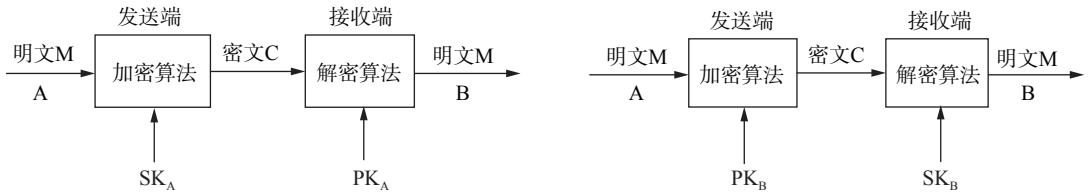


图7.4　非对称密码体制

非对称密码体制以加密算法的函数单向性（即求逆的困难性）为基础，其优点是可以适应网络的开放性要求。与对称密码体制相比，通信双方在通信前无须事先通过安全的“渠道”交换密钥，同时密钥管理简单得多，尤其可以方便地实现数字签名和认证。但是，非对称密钥算法的加密和解密计算过程比较慢，在网络中传输大量分组时，非对称密码体制并不直接用于加密待传输信息，而是用于对称密码体制的密钥加密传输。

7.2.2　密码算法

密码算法是完成加密和解密的运算函数。在加解密过程中，密码算法起着核心作用，决定了加密的方式和复杂程度。密码算法的发展经历了古典密码、对称密码和非对称密码等阶段。古典密码算法包括替代加密、置换加密等；对称密码算法包括DES和AES等；非对称密码算法包括RSA、Rabin、椭圆曲线等；为了识别信息文件是否被篡改，可以用MD5算法生成信息摘要。

7.2.2.1　DES算法

DES（Data Encryption Standard）是IBM公司研究提出的，1977年被美国国家标准局接收并颁布为用于非国家保密机关的数据加密标准。DES属于典型的对称密码算法，它采用56位的密钥将64位的数据加密成64位的密文。由于安全性因素，美国于1998年以后不再将DES作为数据加密标准，但它至今仍然被广泛使用。

DES算法中，明文按64位进行分组，密钥长64位，密钥事实上是56位参与DES运算。其中第8位、16位、24位、32位、40位、48位、56位、64位是校验位，作用是使每个密钥都有奇数个1，并不参与运算。分组后的明文组和56位的密钥按位替代或交换的方法形成密文组（见图7.5）。

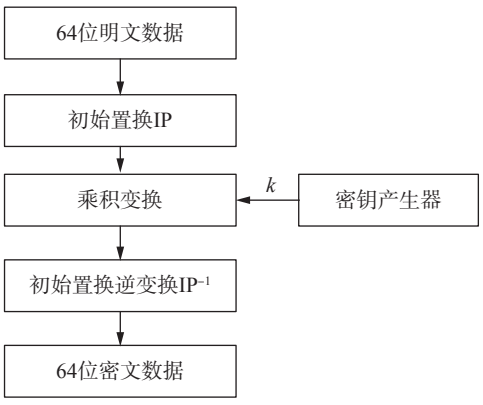


图7.5　DES算法

1. 初始置换IP

输入分组按照初始置换表重排次序，进行初始置换。明文64位，首先经过IP变换。IP代表的是置换规

则表，其功能是把输入的64位数据块按位重新组合，其置换规则如表7.1所示。具体而言，将输入的第58位换到第1位，第50位换到第2位，…，依此类推，最后一位是原来的第7位。

表7.1 IP置换表

58	50	42	34	26	18	10	2
60	52	44	36	28	20	12	4
62	54	46	38	30	22	14	6
64	56	48	40	32	24	16	8
57	49	41	33	25	17	9	1
59	51	43	35	27	19	11	3
61	53	45	37	29	21	13	5
63	55	47	39	31	23	15	7

然后，将初始置换后的结果分成两部分L_0和R_0，L_0是输出的左32位，R_0是右32位。例如，设置换前的输入值为$D_1D_2D_3\cdots\cdots D_{64}$，则经过初始置换后的结果为：$L_0 = D_{58}D_{50}\cdots D_8$；$R_0 = D_{57}D_{49}\cdots D_7$。

2. 乘积变换

DES对经过初始置换的64位明文进行16轮类似的子加密过程。每一轮的子加密过程如下：将64位明文在中间分开，划分为两部分，每部分32位，左半部分记为L，右半部分记为R，以下的操作都是对右半部分数据进行的。

（1）扩展置换。扩展置换将32位的输入数据根据扩展置换表扩展成48位的输出数据。扩展置换的置换方法与初始置换相同，只是置换表不同，扩展置换表如表7.2所示。

表7.2 扩展置换表

32	1	2	3	4	5
4	5	6	7	8	9
8	9	10	11	12	13
12	13	14	15	16	17
16	17	18	19	20	21
20	21	22	23	24	25
24	25	26	27	28	29
28	29	30	31	32	1

（2）异或运算。将48位的明文数据与48位的子密钥进行异或运算。

（3）S盒置换。S盒置换是非线性的，48位输入数据根据S盒置换表置换成32位输出数据。经过异或运算得到的48位输出数据要经过S盒置换，置换由8个盒完成，记为S盒。每个S盒都有6位输入，4位输出（见图7.6）。

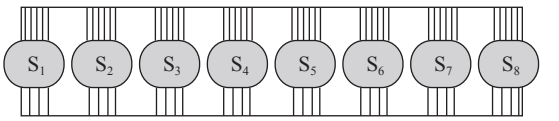


图7.6 S盒置换

这8个S盒是不同的，每个S盒的置换方法如表7.3所示。

表7.3　S盒置换表

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	S1															
0	14	4	13	1	2	15	11	8	3	10	6	12	5	9	0	7
1	0	15	7	4	14	2	13	1	10	6	12	11	9	5	3	8
2	4	1	14	8	13	6	2	11	15	12	9	7	3	10	5	0
3	15	12	8	2	4	9	1	7	5	11	3	14	10	0	6	13
	S2															
0	15	1	8	14	6	11	3	4	9	7	2	13	12	0	5	10
1	3	13	4	7	15	2	8	14	12	0	1	10	6	9	11	5
2	0	14	7	11	10	4	13	1	5	8	12	6	9	3	2	15
3	13	8	10	1	3	15	4	2	11	6	7	12	0	5	14	9
	S3															
0	10	0	9	14	6	3	15	5	1	13	12	7	11	4	2	8
1	13	7	0	9	3	4	6	10	2	8	5	14	12	11	15	1
2	13	6	4	9	8	15	3	0	11	1	2	12	5	10	14	7
3	1	10	13	0	6	9	8	7	4	15	14	3	11	5	2	12
	S4															
0	7	13	14	3	0	6	9	10	1	2	8	5	11	12	4	15
1	13	8	11	5	6	15	0	3	4	7	2	12	1	10	14	9
2	10	6	9	0	12	11	7	13	15	1	3	14	5	2	8	4
3	3	15	0	6	10	1	13	8	9	4	5	11	12	7	2	14
	S5															
0	2	12	4	1	7	10	11	6	8	4	3	15	13	0	14	9
1	14	11	2	12	4	7	13	1	5	0	15	10	3	9	8	6
2	4	2	1	11	10	13	7	8	15	9	12	5	6	3	0	14
3	11	8	12	7	1	14	2	13	6	15	0	9	10	4	5	3
	S6															
0	12	1	10	15	9	2	6	8	0	13	3	4	14	7	5	11
1	10	15	4	2	7	12	9	5	6	1	13	14	0	11	3	8
2	9	14	15	5	2	8	12	3	7	0	4	10	1	13	11	6
3	4	3	2	12	9	5	15	10	11	14	1	7	6	0	8	13
	S7															
0	4	11	2	14	15	0	8	13	3	12	9	7	5	10	6	1
1	13	0	11	7	4	9	1	10	14	3	5	12	2	15	8	6
2	1	4	11	13	12	3	7	14	10	15	6	8	0	5	9	2
3	6	11	13	8	1	4	10	7	9	5	0	15	14	2	3	12
	S8															
0	13	2	8	4	6	15	11	1	10	9	3	14	5	0	12	7
1	1	15	13	8	10	3	7	4	12	5	6	11	0	14	9	2
2	7	11	4	1	9	12	14	2	0	6	10	13	15	3	5	8
3	2	1	14	7	4	10	8	13	15	12	9	0	3	5	6	11

表7.3的使用方法如下：48位的输入分成8组，每组6位，分别进入8个S盒。将每组的6位输入记为$B_0B_1B_2B_3B_4B_5$，那么表中的行号由B_0B_5决定，而列号由$B_1B_2B_3B_4$决定。例如，第一个分组111000要进入第一个S盒S_1，那么行号为10(B_0 B_5)即行号为2，列号为1100(B_1 B_2 B_3 B_4)即列号为12，所对应的数据为3，所以这个S盒的4位输出就是3的二进制表示，即0011。

（4）P盒置换。S盒置换后的32位输出数据根据P盒置换表进行P盒置换，如表7.4所示。P盒置换的过程与初始置换的相同。

表7.4 P盒置换表

16	7	20	21
29	12	28	17
1	15	23	26
5	18	31	10
2	8	24	14
32	27	3	9
19	13	30	6
22	11	4	25

（5）异或操作。经过直接置换的32位输出数据与本轮的*L*部分进行异或操作，结果作为下一轮子加密过程的*R*部分。本轮的*R*部分直接作为下一轮子加密过程的*L*部分。然后进入下一轮子加密过程，直到16轮全部完成。

3. IP逆置换

按照IP逆置换表进行IP逆置换，64位输出就是密文，如表7.5所示。这个置换的过程与初始置换的相同。

表7.5 IP逆置换表

40	8	48	16	56	24	64	32
39	7	47	15	55	23	63	31
38	6	46	14	54	22	62	30
37	5	45	13	53	21	61	29
36	4	44	12	52	20	60	28
35	3	43	11	51	19	59	27
34	2	42	10	50	18	58	26
33	1	41	9	49	17	57	25

4. 密钥的产生

在运算中用到了密钥*K*，*K*为一个64位密钥。但由于其中的第8位、16位、24位、…、64位这8位并未参与DES运算，所以它实际算是一个56位的密钥。因此，需要首先经过置换表的变换，将*K*的位数由64 位变成56位。表7.6所示为置换表内容。

表7.6 64位转56位密钥压缩置换表

57	49	41	33	25	17	9
1	58	50	42	34	26	18
10	2	59	51	43	35	27
19	11	3	60	52	44	36
63	55	47	39	31	23	15
7	62	54	46	38	30	22
14	6	61	53	45	37	29
21	13	5	28	20	12	4

在乘积变换中，每一轮子加密过程所用的子密钥是不同的，且其位数为48位。每个子密钥生成都需要经过循环左移和压缩置换。

将56位密钥分为C_0、D_0两部分，各28位，然后分别进行第1次循环左移，得到C_1、D_1，将C_1(28位)、D_1(28位)合并得到56位，再经过缩小选择换位表，从而得到子密钥K_1(48位)，如表7.7所示。以此类推，便可得到K_2、K_3、…、K_{16}。

表7.7　56位转48位密钥压缩置换表

14	17	11	24	1	5
3	28	15	6	21	10
23	19	12	4	26	8
16	7	27	20	13	2
41	52	31	37	47	55
30	40	51	45	33	48
44	49	39	56	34	53
46	42	50	36	29	32

需要注意的是，16次循环左移对应的左移位数是不同的，每轮的循环左移位数要依据下述规则进行：

1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1

这代表比如第一轮循环左移1位，第二轮循环左移1位，第三轮循环左移2位。

以上是DES算法的过程，该算法过程比较复杂，涉及16轮变换，在16轮变换中，子密钥也是不断变换的。DES加密算法可以用于保存信息时加密，也可以用于通信传输时加密。

7.2.2.2　RSA算法

1976年，Diffie和Hellman为解决密钥管理问题，在他们的奠基性的工作“密码学的新方向”一文中，提出一种密钥交换协议，允许在不安全的媒体上通过通信双方交换信息，安全地传送秘密密钥。在此新思想的基础上，很快出现了非对称密码体制，即公钥密码体制。在公钥体制中，加密密钥不同于解密密钥，加密密钥公之于众，谁都可以使用；解密密钥只有解密人自己知道。它们分别称为公开密钥和私有密钥。在迄今为止的所有公钥密码体系中，RSA系统是最著名并且使用最多的一种。

RSA公钥密码系统是由R. Rivest、A. Shamir和L. Adleman于1977年提出的。RSA的取名就是来自于这三位发明者的姓的第一个字母。

RSA的安全性依赖于大数分解。公开密钥和私有密钥都是两个大素数（大于100个十进制位）的函数。下面描述密钥对是如何产生的。

（1）选择两个大素数p和q，计算$n = p \times q$，$\phi(n) = (p-1) \times (q-1)$。

（2）随机选择与$\phi(n)$互质的数e，要求$e < \phi(n)$。

（3）计算d，满足$e \times d \equiv 1 \bmod \phi(n)$，即$d$与$e$的乘积和1模$\phi(n)$同余。

（4）于是，数(e, n)是加密密钥，(d, n)是解密密钥。两个素数p和q不再需要，应该丢弃，不要让任何人知道。

加密信息m时，首先把m分成等长数据块$m_1, m_2, \cdots, m_i$，块长s，其中$2s \leqslant n$，s尽可能的大。对应的密文

$$c_i = m_i^e \bmod n \tag{a}$$

解密时进行如下计算：

$$m_i = c_i^d \bmod n \tag{b}$$

RSA也可用于数字签名，方案是用上式(a)签名，上式(b)验证。具体操作时，考虑到安全性和m信息量较大等因素，一般是先对m做HASH运算，再进行数字签名。

对于巨大的质数p和q，计算乘积$n = p \times q$非常简便，而逆运算即根据n找到p和q却非常难，这是一种“单向性”，相应的函数称为“单向函数”。任何单向函数都可以作为某一种公钥密码系统的基础，而单向函数的安全性也就是这种公钥密码系统的安全性。

RSA算法安全性的理论基础是，大数的因子分解问题至今没有很好的算法，因而依据公钥e和n不易求出p、q及d。RSA算法要求p和q是两个足够大的素数（例如100位十进制数）且长度相差比较小。

为了说明该算法的工作过程，下面给出一个简单例子。显然，这里只能取很小的数字，但是如上所述，为了保证安全，在实际应用上我们所用的数字要大得多。

例：选取$p = 3$，$q = 11$，则$n = 33$，$\phi(n) = (p - 1) \times (q - 1) = 20$。选取$e = 13$（大于0，且小于$\phi(n)$，并与$\phi(n)$互质，即最大公约数是1），通过$13 \times e \equiv 1 \bmod 20$，计算出$d = 17$。

假定明文为整数$M = 8$，则密文C为

$$\begin{aligned} C &= M^e \bmod n \\ &= 8^{13} \bmod 33 \\ &= 549755813888 \bmod 33 \\ &= 17 \end{aligned}$$

复原明文M为

$$\begin{aligned} M &= C^d \bmod n \\ &= 17^{17} \bmod 33 \\ &= 827240261886336764177 \bmod 33 \\ &= 8 \end{aligned}$$

公钥加密方法既可以用于加密，也可以用于“签名”，以便接收方能确定接收到的信息不是伪造的。

7.2.2.3 MD5算法

MD5（Message Digest 5）是典型的单向散列算法，它主要用于对一段信息产生信息摘要。MD5将一个文件作为一个大文本信息，通过其不可逆的字符串变换算法，产生了文件唯一的MD5摘要。比如，在Windows下有很多软件在下载时都有一个文件名相同、文件扩展名为.md5的文件，在这个文件中通常只有一行文本，大致结构如下：

MD5（readme.txt）= 0ca175b9c0f726a831d895e269332461

上面表示的就是readme.txt文件的信息摘要。MD5是指将整个文件当成一个大文本信息，通过其不可逆的字符串变换算法，产生了这个唯一的MD5信息摘要。

MD5变换首先将输入的信息以512位分组来处理，对每一个512位的分组又划分为16个32位子分组作为输入值。经过MD5算法变换后，算法的输出由4个32位分组组成，将这4个32位分组级联后将生成一个128位散列值，为最后结果。

在处理输入值时，首先需要对信息进行填充，使其位长对512求余的结果等于448。因此，不管信息是多长，信息的位长（Bits Length）都会被扩展至*N**512 + 448（其中，*N*为一个非负整数，可以是零）。填充的方法如下，在信息的后面填充一个1和无数个0，直到满足上面的条件时才停止用0对信息的填充。然后，在这个结果后面附加一个以64位二进制表示的填充前信息长度。经过这两步的处理，现在的信息的位长 = *N**512 + 448 + 64 =(*N* + 1)*512，即长度恰好是512的整数倍。这样做的原因是为满足后面处理中对信息长度的要求。此后，将输入文本的每个512长度的分组再分为32长度的子分组，以用于后续计算。

MD5中有四个32位被称为链接变量（Chaining Variable）的整数参数，它们分别为：*A* = 0x67452301，*B* = 0xefcdab89，*C* = 0x98badcfe，*D* = 0x10325476。

当设置好这四个链接变量后，就开始进入算法的四轮循环运算。循环的次数是信息中512位信息分组的数目，即*N* + 1。

将上面四个链接变量复制到另外四个变量中：*A*到*a*，*B*到*b*，*C*到*c*，*D*到*d*。

主循环有四轮，每轮循环都很相似。第一轮进行16次操作。每次操作对*a*、*b*、*c*和*d*中的其中三个进行一次非线性函数运算，然后将所得结果加上第四个变量、文本的一个子分组和一个常数。再将所得结果向左循环移动一个不定的数，并加上*a*、*b*、*c*或*d*中之一。最后用该结果取代*a*、*b*、*c*或*d*中之一。

以下是每次操作中用到的四个非线性函数（每轮一个）：

$$F(X, Y, Z) = (X \,\&\, Y) \mid ((\sim X) \,\&\, Z)$$
$$G(X, Y, Z) = (X \,\&\, Z) \mid (Y \,\&\, (\sim Z))$$
$$H(X, Y, Z) = X \wedge Y \wedge Z$$
$$I(X, Y, Z) = Y \wedge (X \mid (\sim Z))$$

“&” 代表与操作，“|” 代表或操作，“~” 代表非操作，“^” 代表异或操作。

这四个函数表明：如果*X*、*Y*和*Z*的对应位是独立和均匀的，那么结果的每一位也应是独立和均匀的。

每个人都有自己独特的指纹，而MD5值就像信息串的指纹，信息串的任何变化都会体现在MD5值的变化上。在实际应用中，我们常常会在某些下载站点提供的软件说明信息上看到该软件的MD5值。下载该软件后，可以使用MD5值生成软件计算该软件的MD5值，再与网上提供的标准MD5值进行比对。通过MD5值比对，可以判断文件是否遭受黑客篡改或感染计算机病毒。MD5算法的另一个应用是可以用来作为数字签名。当一方发布一个文件时，计算其MD5值并将这个MD5值提交给第三方公证。如果接收文件方篡改这个文件，则第三方可以通过计算MD5值发现问题。这就是防止抵赖的一种数字签名行为。

7.2.3　密钥管理

在一个密码系统中，按照加密的内容不同，密钥可以分为主密钥、密钥加密密钥和会话密钥。

主密钥（Master Key）是对密钥加密密钥进行加密的密钥，存于主机的处理器中。

密钥加密密钥（Key Encrypting Key，KEK）是传送会话密钥时采用的密钥。

会话密钥是指两个通信终端用户在一次会话或交换数据时所用的密钥。一般由系统通过密钥交换协议动态产生。它使用的时间很短，从而限制了密码分析者攻击时所能得到的同一密钥加密的密文量。丢失时对系统保密性的影响不大。

密钥管理是针对多种不同密钥展开的，密钥管理的内容包括密钥从产生到销毁的各个方面。密钥管理可以设计一套管理体制，还可以通过密钥管理协议对密钥自动管理，内容包括密钥的产生、分配、更换和注入等。对于军用计算机网络系统，由于用户机动性更强，隶属关系和协同作战指挥方式等更复杂，因此对密钥管理提出了更高的要求。

密钥管理的内容覆盖了密钥的整个生命周期：包括密钥生成、分发、有效期与更新、存储、备份和销毁等。下面详细介绍这些内容。

1. 密钥生成

在密钥生成时，有一个规则就是密钥长度应该足够长。一般来说，密钥长度越大，对应的密钥空间就越大，攻击者使用穷举法猜测密钥的难度也就越大。

对于对称加密算法来说，密钥的生成可以由用户自己来设定，也可以用一些软件或硬件设备辅助生成。由自动处理设备生成的随机比特串强度更高。

对于非对称加密算法来说，密钥生成会比较困难，因为密钥必须满足某些数学特征。因此，密钥生成可以通过在线或离线的交互协商方式实现，即利用一些密码协议来实现。

2. 密钥分发

密钥由密钥管理中心或使用者生成后，需要通过一定的方式把密钥传递出去，这就是密钥的分发。当使用对称加密算法进行保密通信时，加密密钥和解密密钥是同一个密钥，发送方需要通过网络把密钥传递给接收方用来解密。当采用非对称加密算法进行通信时，需要使用一对密钥，其中私钥自己保留，公钥要传递给通信对方，这时也需要通过网络进行密钥的传递。

在网络中传输之前，密钥会附加一些检错和纠错位。当密钥在传输中发生错误后，接收方可以通过比较这些检错和纠错位查错。如果发现密钥出错，则接收方请求发送方重传密钥。例如，发送方用密钥加密一个常量，然后把密文的前2~4字节与密钥一起发送。在接收端做同样的工作，如果接收方解密后的常数与发送方的常数相匹配，则传输无错。

3. 密钥有效期与更新

加密密钥不能无限期使用，因为密钥使用时间越长，就越容易泄露。密钥使用越久，越容易被密码专家分析出来。敌对方会尽力搜集更多的明文和密文，密码专家通过明文和密文的规律进行分析，就可能破译出密钥来。时间越长，敌对方搜集的明文和密文越多，就越容易分析出密钥来。敌对方为了窃取密钥，有时还采用穷举攻击法。穷举攻击又称暴力破解，过程是首先把一定长度的字母或数字任意排列组合形成一个密钥字典文件，然后逐个用字典文件中的密钥探测系统的密钥，如果不匹配就测试下一个密钥，直到匹配出密钥。穷举攻击计算需要耗费一定时间，密钥使用时间越长，被破解的可能性越大。如果密钥已泄露但还在使用，那么使用时间越久，损失就越大。

不同的密钥应有不同的有效期。会话密钥的有效期是一次会话过程，在网络中每次会话结束后，发起新会话时，就需要使用新的会话密钥。密钥加密密钥无须频繁更换，因为它们只是偶尔用来作为密钥交换。在某些应用中，密钥加密密钥仅一月或一年更换一次。公开密钥密码应用中的私钥的有效期是根据应用的不同而变化的，有时用来作为数字签名和身份识别的私钥常常持续数年。

旧的密钥到期后，就需要新的密钥代替旧密钥，这就是密钥更新。密钥更新有两种方式：一种是用户以旧密钥作为关键值，并运用密钥生成算法自行产生新的密钥；一种是由密钥中心统一生成并下发新密钥。

4. 密钥存储

密钥可以存储在人脑、文件、磁盘或智能卡中。在保存时，可以把密钥平分成多个部分，由不同的人分管，或者放到不同的存储介质中。为了提高安全性，还可以使用密钥加密密钥对密钥进行加密保存。

5. 密钥备份

最简单的密钥备份办法是使用密钥托管中心。密钥托管要求所有用户将自己的密钥交给密钥托管中心，由密钥托管中心备份保管密钥（如锁在某个地方的保险柜里或用主密钥对它们进行加密保存），一旦用户的密钥丢失（如用户遗忘了密钥或用户意外死亡），按照一定的规章制度，可从密钥托管中心索取该用户的密钥。另一个备份方案是使用智能卡进行临时密钥托管。如管理员A把密钥存入智能卡，当管理员A不在时就把它交给管理员B，管理员B可以利用该卡进行管理员A的工作。当管理员A回来后，管理员B交还该卡，由于密钥存放在卡中，所以管理员B不知道密钥是什么。

6. 密钥销毁

如果密钥必须替换，旧钥就必须销毁。如果密钥写在纸上，那么必须切碎或者烧掉。如果密钥存储在光盘中，光盘应该掰成小块或用专用设备碎掉。如果密钥存储在硬盘中，应重写覆盖磁盘存储的实际位置，当然最好的方式还是物理粉碎。

在密钥管理的六个阶段，其中密钥的生成和密钥的分配是密钥管理使用最为基础和核心的内容。

7.2.4　密码技术应用

密码技术在计算机网络的各层次都得到了应用，比如在数据链路层有链路加密，在网络层有安全协议，在应用层有邮件加密等。下面主要介绍邮件加密软件PGP（Pretty Good Privacy）。

PGP主要是由Philip R. Zimmermann开发的，于1991年在Internet上免费发布。PGP主要针对电子邮件易遭受不明身份者的窃取、篡改、冒用甚至恶意破坏等威胁，提供数字签名、消息加密、压缩、电子邮件兼容性、分段等五种服务，以保证电子邮件的传输安全，如表7.8所示。

表7.8　PGP服务概述

功能	使用的算法	描述
数字签名	DSS/SHA或RSA/SHA	消息的Hash码利用SHA-1产生。将此消息摘要和消息一起用发送方的私钥按DSS或RSA加密
消息加密	CAST或IDEA，或使用3DES或RSA	将消息用发送方生成的一次性会话密钥按CAST-128或IDEA或3DES加密。用接收方公钥按RSA加密会话密钥，与消息一起加密
压缩	ZIP	消息在传输或存储时可用ZIP压缩
电子邮件兼容性	base 64转换	为了对电子邮件应用提供透明性，一个加密消息可以用base 64转换为ASCII码
分段		为了符合最大消息尺寸限制，PGP执行分段和重新组装

PGP并没有采用新的加密算法，而是较好地将对称加密和非对称加密结合起来，综合运用了RD5、RSA及IDEA等算法。PGP加密过程如图7.7所示。

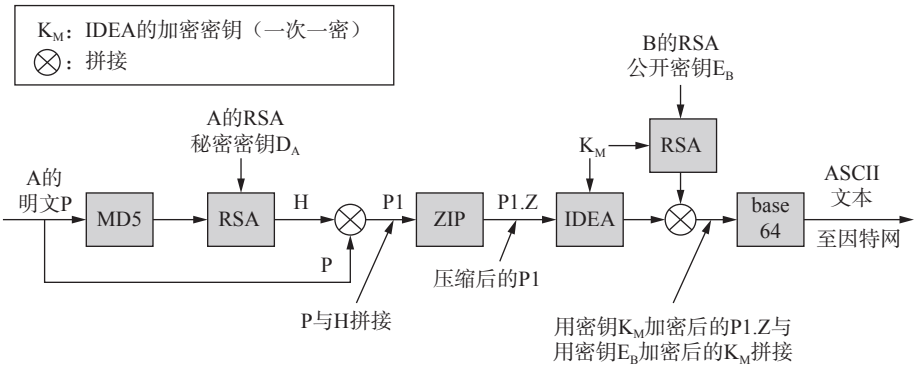


图7.7 PGP加密过程

用户A向用户B发送一个电子邮件明文P，用PGP进行加密。假定用户A和用户B都有RSA的秘密密钥和公开密钥，都有对方的公开密钥。

明文P先经过MD5运算，生成报文摘要；再用RSA的秘密密钥D_A对报文摘要进行加密，得出H；明文P和RSA的输出H拼接在一起，成为另一个报文P1。经ZIP程序压缩后，得出P1.Z；对P1.Z进行IDEA加密（使用的是一次一密的加密密钥，即128比特位的K_M）；密钥K_M再经过RSA加密（使用B的公开密钥E_B）；加密后的K_M与加密后的P1.Z拼接在一起，用base64进行编码；得出ASCII码的文本（只包含52个字母、10个数字和3个符号+，/，=）发送到因特网上。

用户B收到加密的邮件后，先进行base64解码，并用其RSA秘密密钥解出IDEA的密钥。用此密钥恢复出Pl.Z。对Pl.Z进行解压后，还原出P1。B接着分开明文P和加密了的MD5，并用A的公开密钥解出MD5。若与B自己算出的MD5一致，则可认为P是从A发来的邮件。

从加密过程来看，在两个地方使用了RSA：对128比特位的MD5加密和对128比特位的IDEA密钥加密。虽然RSA的运算很慢，但这里只对数量不大的256比特位进行加密。

PGP支持三种RSA密钥长度：384比特位（偶尔使用），512比特位（商业用）和1024比特位（军用）。由于PGP很难被攻破，因此目前可以认为PGP是足够安全的。

7.3 信息隐藏技术

信息隐藏作为一门新兴的交叉学科，伴随着信息和网络技术的飞速发展，在隐蔽通信、数字版权保护等方面起着越来越重要的作用。

7.3.1 基本概念

信息隐藏是指利用载体信息的冗余性，将秘密信息隐藏于普通信息中，通过普通信息的发布而将秘密信息发布出去，即将重要的信息隐藏于其他信息里面从而掩饰它的存在。它隐藏的是信息的“存在性”，使它们看起来与一般非机密资料没有区别，可以避免引起其他人注意，从而具有更大的隐蔽性和安全性，非常容易逃过拦截者的破解。

信息隐藏是上世纪90年代开始兴起的信息安全新技术，并成为信息安全技术研究的热点。传统通信领域为了保证传递的信息不被窃听或破坏，常采用密码来保护信息，即让窃听者无法看到或听懂，但是这种技术的缺点是让窃听者知道这就是秘密信息，特别是随着计算机技术的发展，密码的安全性受到很大挑战。而新的信息隐藏技术是将需要传递的秘密信息，隐藏在一

个普通的非秘密消息中，再进行传输，这样即使窃听者窃听了传输的信息，也只会将其当成普通的消息，而不会怀疑或者无法得知是否有秘密信息的存在。

一般而言，信息隐藏分为四个阶段：预处理阶段、嵌入阶段、传输阶段和提取阶段。为了在每个阶段都保证安全，必须在预处理阶段引入加密算法。在嵌入阶段使用基于小波的隐藏信息的算法。在传输阶段进行隐蔽通信，从而使传输阶段也是安全的。所以，这套信息隐藏的处理方案将形成一个安全的体系，既能隐藏秘密信息的内容，也能隐蔽通信的接收方和发送方，从而建立隐蔽通信。

信息隐藏具有鲁棒性、不可检测性、透明性、安全性和自恢复性等特征。

- 鲁棒性。不因图像文件的某种改动而导致隐藏信息丢失的能力。这里所谓的“改动”包括传输过程中的信道噪声、滤波操作、重采样、有损编码压缩、数模或模数转换等。
- 不可检测性。隐蔽载体与原始载体具有一致的特性。如具有一致的统计噪声分布等，以便使非法拦截者无法判断是否有隐蔽信息。
- 透明性。利用人类的视觉系统或听觉系统的属性，经过一系列隐藏处理，使目标数据没有明显的降质现象，而隐藏的数据却无法被人看见或听见。
- 安全性。隐藏算法有较强的抗攻击能力，即它必须能够承受一定程度的人为攻击，而使隐藏信息不会被破坏。
- 自恢复性。由于经过一些操作或变换后，可能会使隐蔽载体产生较大的破坏，如果只从留下的片段数据仍能恢复隐藏信号，而且恢复过程无须宿主信号，这就是所谓的自恢复性。

7.3.2　基本方法

信息隐藏的方法主要有隐写术、数字水印、可视密码、潜信道和隐匿协议等。

1. 隐写术

隐写术是将秘密信息隐藏在某些宿主对象中，且信息传输或存储过程中不被发现和引起注意，接收者获得隐藏对象后按照约定规则可读取秘密信息的技术。现有的隐写术方法主要有利用高空间频率的图像数据隐藏信息的方法，采用最低有效位方法将信息隐藏到宿主信号中的方法，使用信号的色度隐藏信息的方法，在数字图像的像素亮度的统计模型上隐藏信息的方法，以及Patchwork方法等。

2. 数字水印

信息隐藏的一个重要分支是数字水印，它将一些标识信息（即数字水印）直接嵌入数字载体（包括多媒体、文档、软件等）中，但不影响原载体的使用价值，也不容易被人的知觉系统（如视觉或听觉系统）觉察或注意到。

目前主要有两类数字水印，一类是空间数字水印，另一类是频率数字水印。空间数字水印的典型代表是最低有效位（LSB）算法，其原理是通过修改表示数字图像的颜色或颜色分量的位平面，调整数字图像中对人的感知不重要的像素来表达水印的信息，以达到嵌入水印的目的。频率数字水印的典型代表是扩展频谱算法，其原理是通过时频分析，根据扩展频谱特性将水印扩散到一个广泛的频率范围，使得在任何频率单元的能量都很弱，使他人发现不了。数字水印的插入过程和检测过程如图7.8所示。

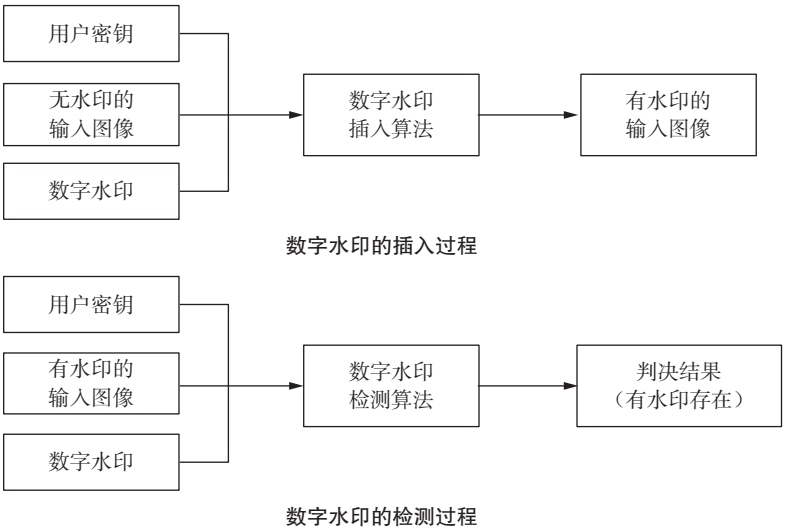


图7.8 数字水印的插入与检测过程

3. 可视密码技术

可视密码技术是Naor和Shamir于1994年首次提出的，其主要特点是恢复秘密图像时无须任何复杂的密码学计算，以人的视觉即可将秘密图像辨别出来。其做法是产生n张不具有任何意义的胶片，任取其中t张胶片叠合在一起即可还原出隐藏在其中的秘密信息。其后，人们又对该方案进行了改进和发展。主要的改进办法有：使产生的n张胶片都具有一定的意义，这样做更具有迷惑性，改进了相关集合的构造方法，将针对黑白图像的可视秘密共享扩展到基于灰度和彩色图像的可视秘密共享。

4. 潜信道

潜信道又称阈下信道。阈下信道的概念是Gustavus J. Simmons于1978年在美国圣地亚国家实验室（Sandia National Labs）提出的，之后又做了大量的研究工作。阈下信道是指在公开（Overt）信道中所建立的一种实现隐蔽通信的信道，这是一种隐蔽（Covert）的信道。绝大多数数字签名方案都可包含阈下信道的通信，其最大特点是阈下信息包含于数字签名中，但对数字签名和验证的过程无任何影响，这正是其隐蔽性所在。阈下信道在国家安全方面的应用价值很大。如果采用全球性标准，那么世界上任何地方的用户检查点都能即时检查出数字证件上的信息完整性，并能确定持证人是否为合法持证人。

7.3.3 信息隐藏的应用

在信息安全领域中，信息隐藏技术的应用可归结为下列几个方面。

1. 数字知识产权保护

数字知识产权保护是信息隐藏技术中数字水印技术和数字指纹技术所力图解决的重要问题，信息隐藏技术的绝大部分研究成果都是在这一应用领域中取得的。随着网络和数字技术的快速普及，通过网络向人们提供的数字服务也会越来越多，如数字图书馆、数字图书出版、数字电视、数字新闻等。这些服务提供的都是数字产品，数字产品具有易修改、易复制、易窃取的特点，因此当前的数字知识产权保护就已经成为迫切需要解决的实际问题。

信息隐藏技术应用于版权保护时，所嵌入的签字信号通常称为“数字水印”，数字水印技术可以成为解决此难题的一种方案。现在越来越多的视频信号、音频信号和数字图像被“贴上”不可见的标签，用以防止非法复制和进行数据跟踪。服务提供商在向用户发送产品的同时，将双方的信息代码以水印的形式隐藏在作品中，这种水印从理论上讲应该是不能被破坏的。当发现数字产品在非法传播时，可以通过提取出的水印代码追查非法散播者。其主要特点是版权保护所需嵌入的数据量较小，对签字信号的安全性和鲁棒性要求很高。

2. 数据完整性鉴定

数据完整性鉴定是指对某一信号的真伪或完整性的判别，并需要进一步指出该信号与原始真实信号的差别，以确认资料在网上传输或存储过程中并没有被篡改、破坏或丢失。假定接收到一个如音频、视频或图像等多媒体信号，并初步判断它很可能是某一原始真实信号的修改版本，数据篡改验证的任务就是在对原始信号的具体内容不可知的情况下，以最大的可能判断其是否为真实的。首先，要充分利用数据库管理系统提供的数据完整性的约束机制和各种输入数据的引用完整性约束设计，以便保证数据完整、准确的输入和存储。其次，在数据传输过程中可视情况选用相应的数据校验方式对传输数据进行校验检查。

3. 数据保密

在网络上传输秘密数据要防止非法用户的截获和使用，这是网络安全的一个重要内容。随着信息技术的发展以及经济的全球化，这一点不仅涉及政治、军事领域，还将涉及商业、金融机密和个人隐私。信息隐藏技术为网上交流的信息采取了有效的保护，比如电子政务中敏感信息、电子商务中的秘密协议和合同、网上银行交易的重要数据、重要文件的数字签名以及个人隐私等，还可以对一些不愿为别人所知的内容使用信息隐藏方式进行隐藏存储，从而使数据得到保密，保证了信息的安全性。

4. 资料不可抵赖性的确认

在网上交易中，交易双方的任何一方不能抵赖自己曾经做出的行为，也不能否认曾经接收到对方的信息，这是交易系统中的一个重要环节。这可以使用信息隐藏技术，在交易体系的任何一方发送和接收信息时，将各自的特征标记形式加入所传递的信息中，这些标记应是不能去除的，从而达到确认其行为的目的。

信息隐藏是一项崭新的技术领域，也是多媒体技术、网络技术研究的前沿，应用前景十分广阔，必将吸引广大图像、语音、网络、人工智能等领域的研究者加入这一行列，从而推动信息安全技术更快地发展。

7.4　网络安全技术

随着计算机网络的发展及普及，计算机网络存在着非常严重的安全风险，面临着越来越多的网络攻击，因此网络安全已成为网络建设中的关键问题。各国都投入了大量的人力、物力来研究网络安全技术，开发各种网络安全产品，使本国的计算机和网络尽可能少受到攻击。

7.4.1　防火墙技术

当前，防火墙应用越来越广泛，已不再是服务器领域的专属，在个人计算机上应用也非常多，因此防火墙在网络安全领域发挥的作用非常大。

7.4.1.1 防火墙的概念

防火墙最初被认为是一个建筑名词。古代构筑和使用木制结构房屋的时候，为防止火灾的发生和蔓延，人们将坚固的石块堆砌在房屋周围作为屏障，这种防护构筑物就称为“防火墙”(Fire Wall)。随着计算机和网络的发展，各种攻击入侵手段也相继出现了，为了保护计算机的安全，人们开发出一种能阻止计算机之间直接通信的技术，并沿用了古代类似这个功能的名字——“防火墙”技术来源于此。

防火墙是位于被保护网络和外部网络之间执行访问控制策略的一个或一组系统，包括硬件和软件，构成一道屏障，以防止发生对被保护网络的不可预测的、潜在破坏性的侵扰。

对于普通用户来说，所谓“防火墙”指的就是一种被放置在自己的计算机与外界网络之间的防御系统，从网络发往计算机的所有数据都要经过它的判断处理后，才会决定能不能把这些数据交给计算机。一旦发现有害数据，防火墙就会将其拦截下来，实现了对计算机的保护功能。

防火墙技术从诞生开始，就在一刻不停地发展着，各种不同结构、不同功能的防火墙构筑成网络上的一道道防御大堤。

7.4.1.2 防火墙主要技术

传统意义上的防火墙技术分为三大类，“包过滤”(Packet Filtering)、“应用代理”(Application Proxy)和“状态监视”(Stateful Inspection)。这三种不同的技术分别工作在OSI网络体系结构的不同层次，它们监视的内容和防护的方式也不相同。无论一个防火墙的实现过程多么复杂，归根结底都是在这三种技术的基础上进行功能扩展的。

1.包过滤技术

包过滤是最早使用的一种防火墙技术，它的第一代模型是“静态包过滤”(Static Packet Filtering)，使用包过滤技术的防火墙通常工作在OSI模型中的网络层(Network Layer)。后来发展了更新的“动态包过滤”(Dynamic Packet Filtering)。

包过滤防火墙可以理解为一台有数据包过滤能力的路由器，它工作在网络层(见图7.9)。通过在防火墙上定义相应的包过滤规则(如表7.9所示的规则实例)，用来匹配数据包内容并决定哪些包应被放过，哪些包应被拒绝。当拒绝数据包时，可以通知发送者丢弃了哪些数据，也可以不发任何通知直接丢弃这些数据。

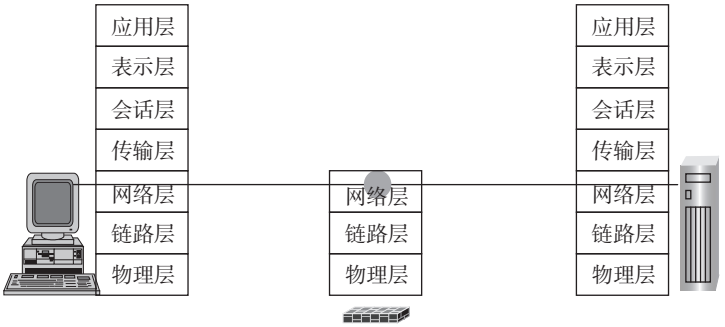


图7.9 包过滤防火墙

包过滤防火墙能对以下信息进行分析：

- 网络层的源和目的地址
- 网络层协议信息

- 传输层协议信息
- 发送或接收数据包的端口

对于没有匹配规则集的情况，用户可以设置包过滤防火墙用以下策略之一处理：

（1）如果没有匹配规则集，则允许这些数据包通过；

（2）如果没有匹配规则集，则丢弃这些数据包。

表7.9　包过滤防火墙规则实例

规则	方向	源地址	目的地址	协议	源端口	目的端口	动作
001	入	任意	172.46.23.45	TCP	/	25	拒绝
002	出	172.46.23.45	任意	TCP	/	>1023	任意
003	出	172.46.23.45	任意	TCP	/	25	允许
004	入	任意	172.46.23.45	TCP	/	>1023	允许

包过滤技术的优点如下。

（1）保护整个网络。包过滤防火墙是两个网络之间访问的唯一通道，所有的通信必须通过它，绕过是困难的。

（2）对用户透明。包过滤不需要用户软件的支持，也不要求对客户机进行特别的设置，也没有必要对用户进行任何培训，因此，很多用户甚至感觉不到包过滤功能的存在，只有在某些包被禁入或禁出时，用户才认识到它与普通路由器的区别。

（3）可使用路由器，不需要其他设备。可以直接在路由器进行配置规则，无须其他专用设备。

包过滤技术的缺点如下。

（1）包过滤的一个重要的局限是它不能分辨好的和坏的用户，只能区分好的包和坏的包。它只能工作于网络层和传输层，并不能判断高级协议里的数据是否有害，比如远程猜测密码等。

（2）包过滤规则难配置。因为包过滤防火墙很复杂，人们经常会忽略建立一些必要的规则，或者错误配置了已有的规则，在防火墙上留下漏洞。

（3）IP欺骗。如果攻击者把自己主机的IP地址设成一个合法主机的IP地址，就可以很轻易地通过包过滤防火墙了。

7.4.1.3　状态监测技术

“状态监测”（Stateful Inspection）技术在保留了对每个数据包的头部、协议、地址、端口、类型等信息进行分析的基础上，进一步发展了“会话过滤”（Session Filtering）功能。在每个连接建立时，防火墙会为这个连接构造一个会话状态，里面包含了这个连接的所有信息，以后该连接都基于这个状态信息进行。这种检测的高明之处是能对每个数据包的内容进行监视，一旦建立了一个会话状态，则此后的数据传输都要以此会话状态作为依据。状态监测防火墙通常监测通信双方TCP和UDP连接的建立情况，因此是主要基于传输层的一种防护方式，在监测连接过程，同时也考虑网络层和高层应用的情况（见图7.10）。

例如，主机A打开一个到网站服务器B的网页连接时，使用一个源端口为5000，目的端口为80的TCP报文，并在控制域中使用了网络请求SYN标记，表示有一个连接请求。当状态防火墙收到这样的流量并且没有过滤规则阻止此流量时，它不像包过滤防火墙那样只简单地允许其通过，而是在其配置中增加一个连接状态信息，这些信息会被添加到已存在的过滤规则集的顶部，也可以添加到一个状态表，这个状态表用来保持跟踪连接的状态。

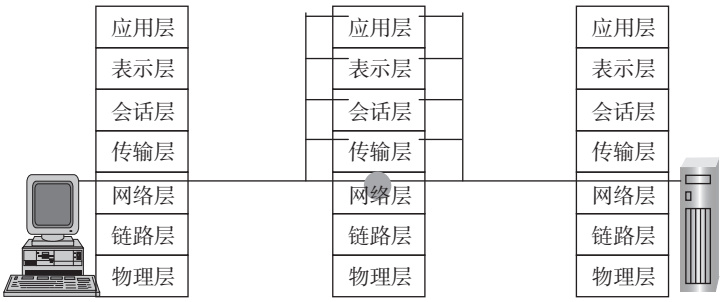


图7.10 传输层进行防护

在B接收到连接请求后，它使用网络响应SYN/ACK标记来响应主机A。当这个报文到达防火墙时，该防火墙首先访问状态表，以查看该连接是否已经存在。防火墙通过查看状态表得知从B的TCP端口80到A的TCP端口5000的响应是已存在连接的一部分，所以允许此流量通过。

状态处理过程的优点在于：当连接终止，源和目的设备拆除连接时，状态监测防火墙可以通过检查TCP控制标记获知此信息，从而动态地将此连接从状态表或过滤规则表里删除。状态监测防火墙比包过滤防火墙更智能，它能理解初始化连接、数据传输及终止连接等连接的各个状态。

7.4.1.4 应用网关技术

应用网关防火墙（Application Gateway firewall，AGF）也称为应用代理防火墙或简称为应用网关。应用网关防火墙是在应用层对计算机提供的安全防护，既能支持一个应用，也能支持有限数量的多个应用的访问验证，这些应用通常包括电子邮件用的简单邮件传输协议（Simple Mail Transfor Protocol，SMTP）、Web服务、域名解析协议（Domain Name Service，DNS）、文件传输协议（File Transfor Protocol，FTP）等（见图7.11）。通过对应用请求和使用应用的用户进行验证，防范了非法的应用行为。

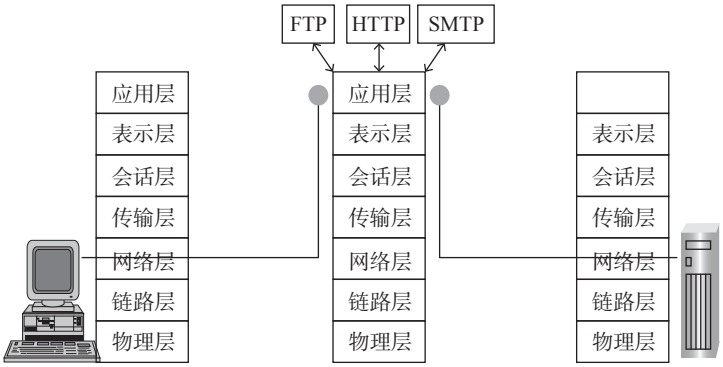


图7.11 应用层进行防护

在应用网关防火墙中，用户可以配置不同应用程序访问网络的权限。比如，可以设置能否从本机访问某网站的网页服务，还可以设置能否从网络中的某台主机向其他网络某主机的FTP上传文件。

以上介绍了防火墙的工作原理和防护技术。防火墙是一种检测通信流量和应用软件，以控制它们能否访问网络的保护手段，它在信息安全保护中扮演着重要的角色。在日常使用中，个人计

算机应该装上防火墙软件，用来防止本机软件随意连接网络，并过滤进出网卡的数据，隔离黑客的攻击行为。在单位网络出口也应该部署防火墙硬件，防止外网黑客对内网的探测和攻击，也可以控制内网中不同网段对其他网段的访问。

但是，防火墙也具有以下局限性。

（1）防火墙不能防范没有经过防火墙的攻击；

（2）防火墙不能解决来自内部网络的攻击和安全问题；

（3）防火墙不能防止策略配置不当或错误配置引起的安全威胁；

（4）防火墙不能防止可接触的人为或自然的破坏；

（5）防火墙不能防范利用标准网络协议缺陷进行的攻击；

（6）防火墙不能防范利用漏洞进行的攻击；

（7）防火墙不能防范数据驱动式的攻击；

（8）防火墙不能防范本身安全漏洞的威胁。

7.4.2　入侵检测技术

对于大部分防火墙来说，它只能对外部的攻击起到防御作用，而无法防御内部的攻击。入侵检测作为一种积极主动的安全防护技术，提供了对内部攻击、外部攻击和误操作的实时保护，在网络系统受到危害之前拦截和响应入侵。

7.4.2.1　入侵检测原理

1. 入侵检测的概念

入侵检测技术是一种主动保护自己的网络和系统免遭非法攻击的网络安全技术。它首先从计算机系统或者网络收集信息并进行分析，然后发现企图破坏计算机资源完整性、机密性和可用性的行为，最后做出相应的反应。

美国国际计算机安全协会（International Computer Security Association，ICSA）对入侵检测的定义是：通过从计算机网络或计算机系统中的若干关键点收集信息并对其进行分析，从中发现网络或系统中是否有违反安全策略的行为和遭到袭击的迹象的一种安全技术。

入侵检测系统（IDS）是防火墙的合理补充，对系统的运行状态进行监视，发现各种攻击企图，以保证系统的安全性。它是防火墙之后的第二道安全防线。

2. 入侵检测系统的功能

入侵检测系统能在入侵攻击对系统发生危害之前将其检测到，并利用报警与防护系统驱逐入侵攻击；在面临入侵攻击时，能减少入侵攻击所造成的损失；在被入侵攻击后，可以收集入侵攻击的相关信息，作为防范系统的知识添加到知识库内，以增强系统的防范能力。入侵检测系统的功能可以具体归为下面几类。

（1）监控、分析用户和系统的活动

这是入侵检测系统能够完成入侵检测任务的前提条件。入侵检测系统通过获取进出某台主机或整个网络的数据，或者通过查看主机日志等信息来实现对用户和系统活动的监控。获取网络数据的方法一般是“抓包”，即将数据流中的所有包截获并进行分析，这就对入侵检测系统的效率提出了较高的要求。如果入侵检测系统不能实时地截获数据包并对它们进行分析，就会出现漏报或网络阻塞的现象。系统的漏报很多，意味着很多攻击行为无法被发现。形成网络阻

塞，就会影响到入侵检测系统所在主机或网络的数据流速，使入侵检测系统成为整个系统的瓶颈。因此，入侵检测系统不仅要能够监控、分析用户和系统的活动，还要能够很快地进行分析。

(2) 发现入侵企图或异常现象

这是入侵检测系统的核心功能，主要包括两个方面：一个作用是入侵检测系统对进出网络或主机的数据流进行监控，看是否有入侵行为。在这种情况下，可能入侵行为正在发生，通过检测攻击行为的迹象，确定是否是攻击行为，从而避免系统遭受攻击。另一个作用是评估系统关键资源和数据文件的完整性，判断系统是否已经遭受了入侵。在这种情况下，攻击的行为已经发生，但可以通过攻击行为留下的痕迹了解攻击行为的一些情况，从而避免再次遭受攻击。而且，对系统资源完整性的检查也有利于对攻击者进行追踪，对攻击行为进行取证。

对于网络数据流的监控，可以使用异常检测的方法，也可以使用滥用检测的方法，这两种检测方法在下文中会详细介绍。检测技术的好坏直接关系到系统能否精确地检测出攻击，因此对于这方面的研究是IDS研究领域的主要工作。

(3) 记录、报警和响应

入侵检测系统是一种主动防御系统。当检测到攻击后，入侵检测系统应该采取相应的措施来记录和阻止攻击，这是所有入侵检测系统必备的功能。在记录、报警和响应功能上，入侵检测系统应该首先记录攻击的基本情况，其次应该能够及时发出报警，最后还应该采取必要的响应行为，如拒绝接收所有来自某台计算机的数据，或者追踪入侵行为等。在记录功能设计方面，好的入侵检测系统不仅能把相关数据记录在文件或数据库中，还应该能提供较强的报表打印功能。此外，实现与防火墙等安全部件的响应互动也是入侵检测系统需要研究和完善的功能之一。

以上是入侵检测系统的基本功能。要实现一个好的入侵检测系统，除了具备以上基本功能外，还可以包括其他一些功能，如审计系统的配置和弱点、评估关键系统和数据文件的完整性等。另外，入侵检测系统应该为管理员和用户提供友好易用的界面，方便管理员设置用户权限、管理数据库、手工设置和修改规则、处理报警，以及浏览和打印数据等。

3. 入侵检测系统的模型

入侵检测系统的模型是构建入侵检测系统的骨架和关键。在入侵检测系统的模型中，设计了多种模块及其工作内容。

(1) CIDF模型

互联网的工程任务组提出的通用入侵检测框架模型（Common Intrusion Detection Framework，CIDF）是当前入侵检测系统的一个通用模型。该系统模型包括四部分内容：事件产生器、事件分析器、事件数据库和响应单元（见图7.12）。

事件产生器的目的是从整个计算环境中获得事件，并向系统的其他部分提供此事件。

事件分析器用于分析事件产生器所产生的事件数据，并生成分析结果。

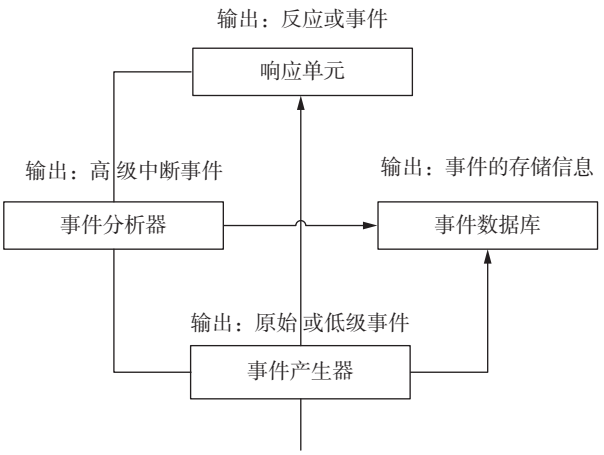


图7.12 CIDF模型

响应单元则是对分析结果做出反应的功能单元，它可以做出切断连接、改变文件属性等强烈反应，也可以只是简单地应答。

事件数据库用于存放各种中间和最终数据，它可以是复杂的数据库，也可以是简单的文本文件。

（2）Denning模型

Denning模型是1978年提出的一个通用入侵检测系统模型（见图7.13）。

在这一模型中，入侵检测系统分析检测主体活动，查看该活动是否会触发规则处理引擎中的规定事件。其中，规则处理引擎是模型中的核心内容。规则处理引擎中既有正常的行为规则及处理方式，也有异常的行为规则及处理方式。对于正常的规则，可以通过新建活动状况建立主体正常行为模型，并通过历史活动状况比对是否为主体正常的行为模型。对于异常行为规则，通过自学习或者特征录入等方式创建异常行为模型。

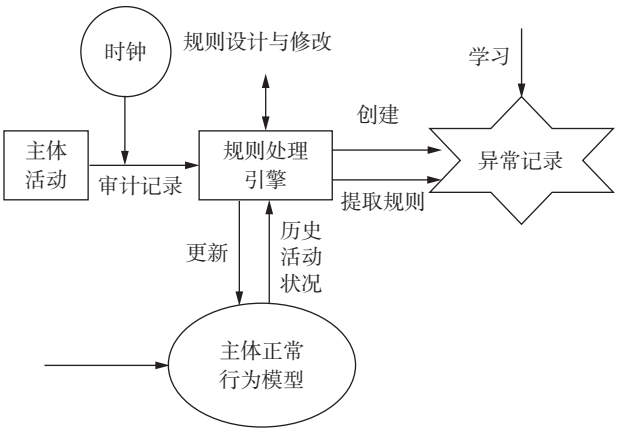


图7.13　Denning模型

7.4.2.2　常用入侵检测技术

从检测技术角度，入侵检测可分为滥用检测和异常检测。

1. 滥用检测技术

滥用检测又称为基于知识的检测，是指运用已知的攻击方法，根据定义好的入侵模式，通过判断这些入侵模式是否出现来进行检测。因为很大一部分的入侵是利用了系统的脆弱性，通过分析入侵过程的特征、条件、排列以及事件之间关系，就能具体描述入侵行为的迹象。这种方法由于依据具体特征库进行判断，所以检测准确度很高，并且因为检测结果有明确的参照，也为系统管理员做出相应措施提供了方便。

常用的滥用检测技术有：简单模式匹配、专家系统、模型推理和状态转换分析。

（1）简单模式匹配。简单模式匹配是最通用的滥用检测技术，主要应用于基于网络的入侵检测系统中，将数据包内容和攻击特征进行匹配比较，若匹配成功则认为是攻击。简单模式匹配的检测原理简单、检测效率高、扩展性好、可以实时检测，但只能适用于比较简单的攻击方式，并且误报率较高。虽然这种技术在检测性能上存在很大的问题，但由于其在系统实现、配置、维护等方面比较方便，因此得到了广泛的应用。

（2）专家系统

专家系统是滥用检测技术中运用得最多的一种方法。专家系统将有关入侵的知识转化为if-then结构的规则，即将构成入侵所要求的条件转化为if部分，将发现入侵后采取的相应措施转化成then部分。if-then结构构成了描述具体攻击的规则库。在判断入侵行为时，专家系统根据审计事件得到状态行为和语义环境，当其中某个或某部分的条件满足时，系统就判断为入侵行为发生。这一工作由推理机自动完成。

(3) 模型推理

模型推理是指结合攻击脚本模型来推测入侵行为是否出现。在模型推理中，首先将有关攻击者行为的经验知识定义为攻击脚本库。攻击脚本库中的经验知识包括：定义攻击者目的、攻击者达到此目的的可能行为步骤，以及对系统的特殊使用等。根据这些知识建立攻击脚本库，每一脚本都由一系列攻击行为组成。然后，在检测时先将攻击脚本的子集看成系统正面临的攻击，通过一个称为预测器的程序模块，根据当前行为模式产生下一个需要验证的攻击脚本子集，并将它传给决策器。决策器在收到信息后，根据这些假设的攻击行为在审计记录中的可能出现方式，将它们翻译成与特定系统匹配的审计记录格式，接着在审计记录中寻找相应的信息来确认或否认这些攻击。设置的初始攻击脚本子集应该易于在审计记录中识别，并且出现频率很高。随着一些脚本被确认的次数增多，另一些脚本被确认的次数减少，攻击脚本不断得到更新。

(4) 状态转换分析

状态转换分析最早由R. Kemmerer提出，即将状态转换法应用于入侵行为的分析。状态转换分析将入侵过程看成一个行为序列，这个行为序列导致系统从初始状态转入被入侵状态。进行分析时，首先针对每一种入侵方法确定系统的初始状态和被入侵状态，以及导致状态转换的转换条件，即导致系统进入被入侵状态必须执行的操作（特征事件）。然后用状态转换表来表示每一个状态和特征事件，这些事件被集成于模型中，所以检测时无须逐个查找审计记录。但是，状态转换是针对事件序列分析的，所以不善于分析过于复杂的事件，而且不能检测与系统状态无关的入侵。

2. 异常检测技术

异常检测（Anomaly Detection）又称为基于行为的检测，是指根据使用者的行为或资源使用状况来判断是否入侵，而不依赖于具体行为特征是否出现来进行检测。异常检测需要首先建立正常用户行为特征轮廓，然后将实际用户行为和这些轮廓进行比较，并标识正常的偏离。如果偏离超过了一定的阈值，则认为发生了入侵。这种检测与具体系统的关系不大，通用性较强。它甚至有可能检测出以前未出现过的攻击方法，不像基于知识的检测那样受已知脆弱性的限制。

常用的异常检测技术有概率统计分析、预测模式生成和神经网络等。

(1) 统计分析。统计分析技术是最早的异常检测技术，首先为每一位系统用户建立历史统计模式，以表示用户正常行为，通过对当前行为与历史统计模式进行比较，来判断是否超过某一阈值，如超过则为入侵行为。比如，CPU占用情况、I/O占用情况、系统调用情况等。其中，所建立的统计模式被定期更新，以便及时反映出用户行为随时间推移而产生的变化。在统计分析中常用的测量参数包括审计事件的数量、间隔时间、资源消耗情况等。目前提出了5种可用于入侵检测的统计模型，包括操作模型、方差、多元模型、马尔可夫过程模型、时间序列分析等。

(2) 预测模式生成技术。该技术试图基于已经发生的事件来预测未来事件，比如存在规则：E1→E2→（E3=80%，E4=15%，E5=5%），即假定事件E1和E2已经发生，E3随后发生的概率是80%，E4随后发生的概率是15%，E5随后发生的概率是5%。如果发生一个与预测统计概率偏差较大的事件，则被标志为攻击。预测模式生成技术的问题在于未被这些规则描述的入侵脚本将不会被标志为入侵。

(3) 神经网络。其想法是用给定的n个动作训练神经网络，以预测用户的下一个命令。训练周期之后，神经网络使用已出现在网中的用户特征匹配实际的用户命令，将统计差异较大的事件标记为非法。使用神经网络的优点是：可以很好地处理噪声数据，因为其只与用户行为相关，而不依赖于任何低层数据特性的统计。

(4) 其他异常检测技术。目前出现了一些新的检测技术，可以将其归入异常检测，如免疫系统、遗传算法等方法，这些方法将生物技术应用到入侵检测系统中，以提高入侵检测能力。

7.4.3　身份认证技术

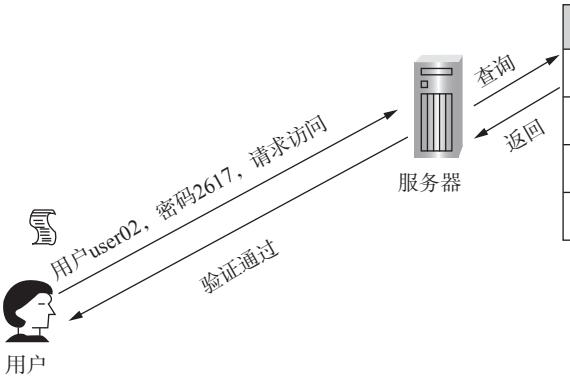
在网络信息传输中，为了确定通信双方的身份，保证发送信息的不可抵赖性，需要用到身份认证。同时，身份认证也是网络中的用户接入某一个系统从而获取信息之前要做的事情，保证了用户身份的合法性和网络的分级授权管理。身份认证可以用多种方式实现，其中口令认证和数字证书是两种常用的技术。

7.4.3.1　口令认证技术

基于口令的认证是指通过用户输入的用户名和口令来确认用户身份的一种机制。基于口令的身份认证是最常见也最简单的一种身份认证机制，例如，电子邮箱、论坛账号等都通过口令来确认对方的身份。同一个用户可能会有多个用户名和口令的组合，但是不同的用户不能有相同的用户名。

1. 静态口令认证

用户设置一个口令信息作为自己的身份登录所使用的密码，系统保存每个用户的一组信息(ID，Password)，即用户的身份信息和口令。当被认证对象要求访问提供服务的系统时，提供服务的认证方要求被认证对象提交口令信息。认证方收到口令后，将其与系统中存储的用户口令进行比较，以确认被认证对象是否为合法的访问者，这种认证方式称为静态口令认证（见图7.14)。



用户名	密码	权限
admin	12xw35U	rw
user01	234xw32	w
user02	2617	r
user03	234123	r

图7.14　静态口令认证

静态口令是目前应用最广泛的认证方法。例如，操作系统、邮件系统等一些应用系统的登录和权限管理大多采用静态口令认证。其优势在于实现的简单性，无须任何附加设备，成本低、速度快。但基于口令的认证方式存在如下安全问题。

(1) 它是一种单因素的认证方式，安全性依赖于口令，口令一旦被泄露，用户即可被冒充。

(2) 用户往往选择简单、容易被猜测的口令。

(3) 口令在传输过程中可能被截获。

(4) 系统中所有用户的口令以文件形式存储在认证方，攻击者可以利用系统中存在的漏洞获取系统的口令文件。即使口令经过加密后存放在口令文件中，如果窃取了口令文件，就可以进行离线的字典式攻击。一旦攻击者能够访问口令表，整个系统的安全性就受到了威胁。

(5) 用户在访问多个不同安全级别的系统时，都要求用户提供口令，用户为了记忆的方便，往往采用相同的口令。而低安全级别系统的口令更容易被攻击者获得，从而用来对高安全级别系统进行攻击。比如，在Windows操作系统中，黑客往往先获取guest用户口令，再通过一些DOS命令把自己加入administrator用户组。

(6) 只能进行单向认证，即系统可认证用户，而用户无法对系统进行认证，攻击者可能伪装成系统骗取用户的口令。

为了有效地改进口令认证的安全性，安全专家提出了一次性口令密码技术，即下面要讲的动态口令认证技术。

2. 动态口令认证

动态口令认证是指根据专门的算法生成不可预测的随机数字组合作为密码，每个密码只能使用一次，目前被广泛应用在网银、网游、电信运营商、电子商务、企业运营等领域。

动态口令可以加载到不同的硬件设备上使用。当前，最常见的动态口令有短信密码、硬件令牌、手机令牌等形式。

(1) 短信密码

短信密码以手机短信形式请求包含6位或更多随机数的动态口令。身份认证系统以短信形式将动态口令发送到客户的手机上，客户在登录或者交易认证时输入此动态口令，从而确保系统身份认证的安全性。

(2) 硬件令牌

当前最主流的是基于时间同步的硬件令牌，动态口令每60 s变换一次，一次有效，产生6位到8位动态数字。

(3) 手机令牌

手机令牌是一种手机客户端软件，基于时间同步方式，每隔60 s产生一个随机6位动态密码，口令生成过程不产生通信及费用，具有使用简单、安全性高、低成本、无须携带额外设备、容易获取、无物流等优点。

7.4.3.2 数字证书技术

数字证书类似于现实中的护照，是人们在网络中的通行证，人们可以在网上用它来识别对方的身份。数字证书是一种权威性的电子文档，由权威公正的第三方机构签发。数字证书一般包括证书版本、序列号、用户标识符、用户的公钥、证书所用的数字签名算法说明等内容。

在实际应用中，往往结合使用数字证书和加密技术，对网络上传输的信息进行加密和解密、数字签名和签名验证，以确保网上传递信息的机密性、完整性。使用了数字证书后，即使用户发送的信息在网上被他人截获，甚至丢失了个人的账户、密码等信息，仍可以保证账户和资金的安全。

数字证书采用的是公钥体制，而当前网络普遍应用的基于公钥体制的认证系统，称为公钥基础设施（Public Key Infrastructure，PKI）。PKI包括：用于用户注册和接受用户的证书请求的注册机构（Registration Authority，RA）；负责生成密钥、发放和管理证书的权威机构（Cerificate Authority，CA）；存储和管理密钥、证书及废止证书列表的数据库及目录服务器；在密钥丢失时，自动恢复密钥、恢复数据的服务器；实现对整个PKI系统的控制和管理的应用模块。有时，CA和RA是同一个单位，同时担负CA和RA的工作，称为认证中心。因此，我们可以理解CA是PKI的核心部分，CA认证的依据是数字证书。

1. 数字证书的申请与颁发

常见的数字证书有三种：域名型证书、企业型证书和增强型证书。域名型证书用于网站身份认证；企业型证书用于企业身份认证；增强型证书是包括网站和企业身份双重认证的一种证书，是具有更高安全级别的证书。这些不同类别的证书的申请既有一些通用步骤，也有独特的审核验证方式。

申请数字证书时，用户一般要携带有关证件到各地的证书受理点，或者直接到证书发放机构（即CA中心）填写申请表并进行身份审核，审核通过后交纳一定费用就可以得到装有证书的相关介质（如磁盘）和一个写有密码口令的密码信封。

申请域名型证书时无须递交书面审查资料，仅需进行域名有效性验证，即可网上申请。企业型证书需要进行严格的网站所有权的真实身份验证，证书标示了企业组织机构的详细情况，以强化信任度。增强型证书除了进行严格的网站所有权的真实身份验证之外，还加入第三方验证，证书标示增强组织机构详情，以强化信任度。

如图7.15所示，数字证书的申请和颁发一般包括如下几个步骤。用户首先产生自己的密钥对，并将公共密钥及部分个人身份信息传送给认证中心。认证中心在核实身份后，将执行一些必要的过程，以确认请求确实由用户发送而来，然后，认证中心将发给用户一个数字证书，该证书内包含用户的个人信息和公钥信息，同时还附有认证中心的签名信息，然后用户就能使用自己的数字证书进行相关的各种活动。数字证书由独立的证书发行机构发布。数字证书各不相同，每种证书可提供不同级别的可信度。

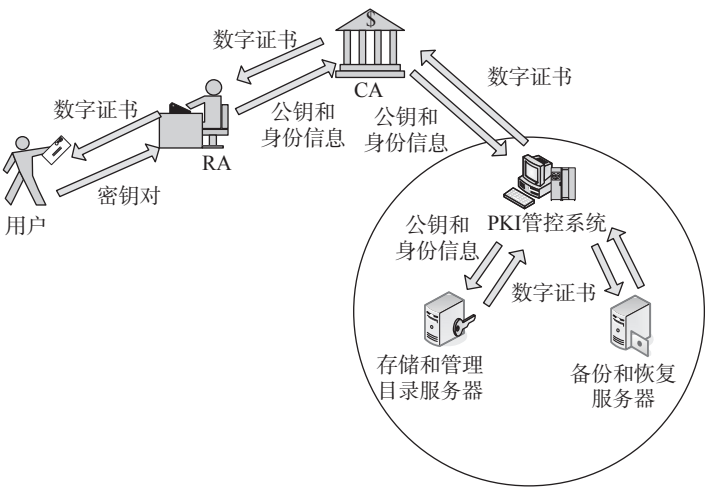


图7.15　数字证书的申请和颁发过程

2. 数字证书的使用及原理

用户在进行需要使用证书的网上操作时，必须准备好装有证书的存储介质。如果用户是在自己的计算机上进行操作，操作前必须先安装CA根证书。所访问的系统如果需要使用数字证书，则一般会自动弹出提示框，要求安装根证书，用户直接选择确认即可；当然也可以直接登录CA中心的网站，下载安装根证书。操作时，一般系统会自动提示用户出示数字证书或者插入证书介质（IC卡或Key），用户插入证书介质后系统将要求用户输入密码，此时用户需要输入申请证书时获得的密码信封中的密码，密码验证正确后系统将自动调用数字证书进行相关操作。使用后，用户应记住取出证书介质并妥善保管。当然，不同系统的数字证书会有不同的使用方式，但系统一般会有明确提示，用户使用起来都较为方便。

数字证书采用公钥密码体制，常用的一种是RSA密码。在数字证书认证过程中，每位用户使用私有密钥（私钥）进行解密和签名；使用公开密钥（公钥）进行加密和验证签名。

如图7.16所示，数字证书认证的具体过程如下。

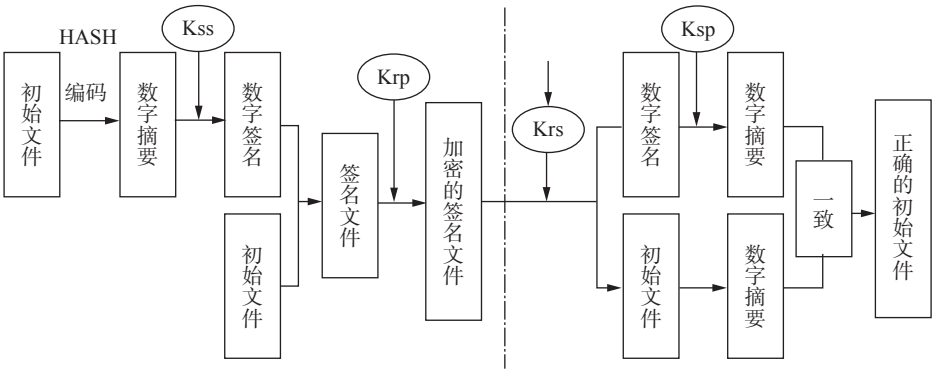


图7.16 数字证书认证过程

(1) 将报文按双方约定的散列摘要算法计算得到一个固定位数的报文摘要。在数学上保证：只要改动报文中的任何一位，重新计算出的报文摘要值就会与原先的值不相符。这样就保证了报文的不可更改性。

(2) 将该报文摘要值用发送者的私人密钥（Kss）加密，而产生的报文即称数字签名。然后连同原报文组合成拟发送的报文。

(3) 当发送一份保密文件时，发送方使用接收方的公钥（Krp）对数据进行加密，这样信息就可以安全无误地到达目的地了。

(4) 接收方则使用自己的私钥（Krs）解密，还原出原报文和数字签名摘要。

(5) 接收方用发送方的公钥（Ksp）解密摘要部分，进行比对，看报文信息是否被篡改。

有时，数字证书认证过程中的数字签名是使用CA的密钥来完成的。CA为某一用户A颁发证书，并用自己的私钥对证书签名。另一用户想验证A的身份时，利用CA的公钥验证A的数字证书的完整性，从而判断A是否是所声称的用户。

数字证书的使用基于公钥加密算法。公钥加密算法可以用来进行通信双方的身份认证。为了验证接收信息与发送信息是否一致，对信息生成报文摘要。这时，发送方用私钥加密的不是信息而是报文摘要。为了保证安全地传输信息，在传输过程中又引入了发送方使用接收方公钥加密的非对称加密。这就是数字证书验证通信双方身份的核心内容。

7.4.4 安全协议

协议是网络中参与通信者互相协商，约定如何进行通信的标准与规则。而安全协议是融合了加密技术和认证技术的安全通信标准，它能够提供安全服务，是保证网络安全的基础。在安全协议中，较为著名的是在网络体系结构模型的网络层和应用层之间启用的SSL协议，以及在网络层上用于建立安全隧道的IPSec协议。

7.4.4.1 SSL协议

安全套接层协议（Secure Socket Layer，SSL）是基于公钥密码体制的安全网络通信协议。SSL协议是网景公司（Netscape）推出的基于网络应用的安全协议，指定了一种在应用程序协议（如网页访问所用的HTTP协议、文件传输所用的FTP协议等）和网络层协议之间提供数据安全性分层的机制。它为网络连接提供数据加密、服务器认证、消息完整性以及可选的客户机认证等服务，主要用于提高应用程序之间数据的安全性。

SSL结合了对称密码技术和公钥密码技术，可以实现如下三个通信目标。

（1）秘密性。SSL客户机和服务器之间传送的数据都经过了加密处理，网络中的非法窃听者所获取的信息都将是无意义的密文信息。

（2）完整性。发送方发送数据之前，SSL利用某种密码算法和散列函数提取所传输信息的特征值。接收方接收数据之后，再根据该种密码算法和散列函数计算信息特征值，与发送来的数值进行比较，查看数据是否被破坏。如果发现数据被破坏，则请求发送方重传，以确保收到的数据是完整无误的。

（3）认证性。利用证书技术和可信的第三方认证，可以让客户机和服务器相互识别对方的身份。为了验证证书持有者是其合法用户（而不是冒名用户），SSL要求证书持有者在握手时相互交换数字证书，通过验证来保证对方身份的合法性。

1. SSL协议的体系结构

SSL协议位于网络体系结构的网络层和应用层之间，使用传输控制协议（Transimission Control Protocol，TCP）来提供一种可靠的端到端的安全服务，它使客户机和服务器应用之间的通信不被攻击窃听，并且始终对服务器进行认证，还可以选择对客户机进行认证。SSL协议在应用层通信之前就已经完成加密算法、通信密钥的协商以及服务器认证工作，在此之后，应用层协议所传送的数据都被加密。SSL由共同工作的两层协议组成（见图7.17）。从体系结构图可以看出SSL安全协议实际是SSL握手协议、SSL修改密文协议、SSL报警协议和SSL记录协议组成的一个协议族。

<table>
<tr><td colspan="3">应用层</td></tr>
<tr><td>SSL 握手协议</td><td>SSL 修改密文协议</td><td>SSL 报警协议</td></tr>
<tr><td colspan="3">SSL 记录协议</td></tr>
<tr><td colspan="3">传输层（TCP 协议）</td></tr>
<tr><td colspan="3">网络层（IP 协议）</td></tr>
</table>

图7.17　SSL体系结构

SSL记录协议为SSL连接提供了两种服务：（1）机密性；（2）消息完整性。当发送用户进行网络操作，从应用层产生待发送数据以后，SSL记录协议首先接收传输的应用报文，将数据分成可管理的块，进行数据压缩（可选），接着利用DES、3DES或其他加密算法进行数据加密，最后增加由内容类型、主要版本、次要版本和数据长度组成的首部。经过处理的数据再经过传输层、网络层等，一层层封装生成物理线路传输的信息，并传递给对方用户。对方用户接收数据之后，从下层向上层传递，经过SSL协议处理时，执行发送数据时SSL协议工作的逆过程，依次解密、验证、解压缩和重新装配数据，然后交给接收用户。

SSL修改密文协议是SSL高层协议中最简单的一个。协议由单个消息组成，该消息只包含一个值为1的单字节。该消息用于协商当前连接使用的密码组。为了保障SSL传输过程的安全性，双方应该每隔一段时间就改变一次加密规范。

SSL报警协议用来为通信双方传递SSL的相关警告。如果在通信过程中某一方发现任何异常，就需要给对方发送一条警示消息。警示消息有两种：一种是 Fatal错误，若传递数据过程中发现错误的物理地址信息，双方就需要立即中断会话，同时消除自己缓冲区相应的会话记录；第二种是Warning消息，这时通信双方通常都只是记录日志，而对通信过程不造成任何影响。

SSL握手协议使服务器和客户机能够相互鉴别对方，使通信双方可以在交换数据前协商具体的加密算法和保密密钥，为下一步记录协议要使用的密钥信息进行协商，使客户机和服务器建立并保持安全通信的状态信息，以保护在SSL记录中发送的数据。SSL握手协议在任何应用程

序数据传输之前使用，包含四个阶段，其中第一个阶段用于建立安全能力；第二个阶段用于服务器鉴别和密钥交换；第三个阶段用于客户鉴别和密钥交换；第四个阶段用于完成握手协议。

2. SSL协议的实现

基于SSL的程序可分为两个部分：客户机和服务器，使用SSL协议使通信双方可以相互验证对方身份的真实性，并且能够保证数据的完整性和机密性。建立SSL通信的过程如图7.18所示。

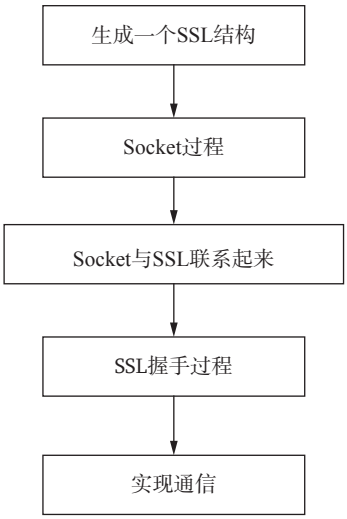


图7.18 SSL通信过程

SSL协议的具体过程如下。

（1）客户机浏览器发送一个连接请求给服务器。

（2）服务器将自己的证书，以及与证书相关的信息发送给客户浏览器。

（3）客户机浏览器检查服务器发送过来的证书是否是自己信赖的CA中心所签发的。如果是，就继续执行协议；如果不是，客户机浏览器就给客户机发送一条警示消息：警示客户机这个证书是不可信赖的，询问客户机是否需要继续。

（4）接着客户机浏览器比较证书里的消息，例如域名和公钥，与服务器刚刚发送的相关消息是否一致，如果是一致的，客户机浏览器认可这个服务器的合法身份。

（5）服务器要求客户机发送客户机自己的证书，收到后，服务器验证客户机的证书，如果没有通过验证，则拒绝连接；如果通过验证，则服务器获得客户机的公钥。

（6）客户机浏览器告诉服务器自己所能够支持的通信对称密码方案。

（7）服务器从客户机发送过来的密码方案中，选择一种加密程度最高的密码方案，用客户机的公钥加密后通知浏览器。

（8）浏览器针对这个密码方案，选择一个通话密钥，接着用服务器的公钥加密后发送给服务器。

（9）服务器接收到浏览器送过来的消息，用自己的私钥解密，获得通话密钥。

（10）服务器、浏览器接下来的通信都采用对称密码方案，对称密钥是加密过的。

SSL协议中融合了公钥认证体系及证书应用，可以保障客户机浏览器到服务器之间的安全的加密访问及身份验证。当在浏览器中输入https://访问网站时，实际上SSL协议已经开始默默工作了。

7.4.4.2 IPSec协议

互联网安全协议（Internet Protocol Security，IPSec）是一个用于保证通过IP网络进行安全的秘密通信的开放式标准框架，在IETF制定的标准的基础上，保证通过公共IP网络的数据通信的保密性、完整性和真实性。IPSec协议不是一个单独的协议，它给出了应用于IP层上以保证网络数据安全的一整套体系结构，包括网络认证（Authentication Header，AH）协议、封装安全载荷（Encapsulating Security Payload，ESP）协议、密钥交换（Internet Key Exchange，IKE）协议和用于网络认证及加密的一些算法等。这些协议用于提供数据认证、数据完整性和保密等保护形式。AH和ESP都可以提供认证服务，但AH提供的认证服务要强于ESP。而IKE主要是对密钥进行交换管理，对算法、协议和密钥这三个方面进行协商。

IPSec规定了如何在对等层之间选择安全协议、确定安全算法和密钥交换，向上提供了访问控制、数据源认证、数据加密等网络安全服务。

1. AH协议

AH协议也称为认证头协议，用于保证IP数据报的完整性和真实性，防止黑客截断数据包或向网络中插入伪造的数据包。AH协议提供无连接的完整性、数据源认证和抗重放保护服务。

图7.19展示了AH协议的格式。各个字段的含义如下所示。

<table>
<tr><td>0</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td></tr>
<tr><td>下一个头</td><td>载荷长度</td><td colspan="2">保留未用</td></tr>
<tr><td colspan="4">安全参数索引（SPI）</td></tr>
<tr><td colspan="4">序列号</td></tr>
<tr><td colspan="4">认证数据（可变长度）</td></tr>
</table>

图7.19　AH协议格式

（1）下一个头：标识被传送数据所属的协议。

（2）载荷长度：认证头包的大小。

（3）保留未用：为将来的应用保留（目前都置为0）。

（4）安全参数索引：与IP地址一同用来标识安全参数。

（5）序列号：单调递增的数值，用来防止重放攻击。

（6）认证数据：包含了认证当前包所必须的数据。

AH报头插在网络层IP协议报头之后，应用层协议报头之前。一般AH为整个数据包提供完整性检查，但如果IP报头中包含“生存期（Time To Live）”或“服务类型（Type of Service）”等值可变的字段，则在进行完整性检查时应将这些值可变的字段去除。

2. ESP协议

前面介绍了AH协议，它可以保护通信免受篡改，但没有对数据进行过变换，数据对于黑客而言仍然是清晰的。为了有效地保证数据传输安全，在IPv6 中有另外一个报头ESP，进一步提供了数据保密性并防止篡改。IPSec协议组设计了安全协议ESP（见图7.20），通过加密而实现保护数据的机密性和完整性。

<table>
<tr><td></td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td><td>0 1 2 3 4 5 6 7</td></tr>
<tr><td colspan="4">安全参数序列</td></tr>
<tr><td colspan="4">序列号</td></tr>
<tr><td colspan="4">载荷（可变长度）</td></tr>
<tr><td>载荷（可变长度）</td><td colspan="3">填充（0~255 字节）</td></tr>
<tr><td colspan="2">填充（0~255 字节）</td><td>填充长度</td><td>下一个头</td></tr>
<tr><td colspan="4">认证数据（可变长度）</td></tr>
</table>

图7.20　ESP协议

ESP报文各字段的含义如下。

（1）安全参数序列：与IP地址一同用来标识安全参数。

（2）序列号：单调递增的数值，用来防止重放攻击。

（3）载荷：实际要传输的数据。

（4）填充：某些块加密算法用此将数据填充至块的长度。

（5）填充长度：以位为单位的填充数据的长度。

（6）下一个头：标识被传送数据所属的协议。

（7）认证数据：包含了认证当前包所必须的数据。

ESP报头的位置在IP报头之后，TCP协议、用户数据报协议（User Datagram Protocol，UDP）等传输层协议报头之前。在IPSec协议嵌套使用的情况下，如果已经有其他IPSec协议在使用，则ESP报头应插在其他任何IPSec协议报头之前。ESP认证报尾的完整性检查部分包括ESP报头、传输层协议报头，应用数据和ESP报尾，但不包括IP报头，因此ESP不能保证IP报头不被篡改。ESP加密部分包括上层传输协议信息、数据和ESP报尾。

以上介绍了SSL和IPSec两种安全协议。这两种安全协议定义了安全通信的数据格式、通信双方传递信息的交换方式，为构造安全通信环境打下了基础。这两种安全协议也常常用于构造虚拟专用网络，即VPN。VPN是为互联网中希望能享受专线的隐私性和安全性的用户提供的一种便利的解决方案。它的核心是利用这些安全协议实现通信隧道的加密，实现通信双方身份的验证。

思考题

1. 什么是信息安全？信息安全的基本属性有哪些？
2. 信息安全面临哪些威胁？
3. 网络信息系统在设计过程中应该遵循哪些原则？
4. 对称密码体制和非对称密码体制的区别是什么？
5. 请描述DES数据加密算法的流程。
6. 在一个RSA公钥密码体制中，已知质数$p=7$，$q=17$，选择公钥$e=5$，$m=19$，计算私钥d，并对明文进行加、解密计算。
7. 报文的保密性与完整性有何区别？什么是MD5?
8. 信息隐藏技术与加密技术的区别是什么？
9. 防火墙的主要功能有哪些？
10. 为什么防火墙不能防范内部攻击？
11. 入侵检测技术主要有哪些？有什么不同？
12. 请简述数字证书的认证过程。
13. 试简述SSL的工作过程。

第8章　典型军事信息系统

军事信息系统已成为信息化战争中重要的军事基础设施，其水平的高低已成为影响国家军事实力和军队整体作战能力的重要因素。最典型的军事信息系统是指挥自动化系统，在世界各军的指挥自动化系统中，又以美军的C^4ISR系统最为先进。所以，本章将主要介绍美军的系统，读者可从中体会军事信息技术的应用对战斗力的重要影响。

8.1　指挥自动化系统概述

指挥自动化系统经历了一个由简单到复杂的发展过程，开始指的是指挥、控制、通信及情报（C^3I）系统，后来增加了计算（Computer），成为C^4I系统，以后又增加了监视及侦察，成为C^4ISR系统。

《中国大百科全书（军事卷）》中对指挥自动化做出了这样的解释："指挥自动化是在指挥系统中，运用以电子计算机为核心的自动化设备和软件系统，使指挥员和指挥机构对所属部队的作战和其他行动的指挥，实现快速和优化处理的措施。其目的是提高军队的指挥效能，最大限度发挥部队的战斗力。而指挥自动化系统作为指挥自动化手段的技术实现，是在现代作战理论指导下，综合运用现代电子信息技术和设备，与作战指挥人员紧密结合，实现对部队和武器实施指挥与控制的人机一体化信息系统。"

无论是美军的指挥自动化系统还是我军的指挥自动化系统，其主要组成部分都包括情报预警系统、军事通信系统、指挥控制系统及火力控制系统等（见图8.1）。如果我们形象地把一个指挥自动化系统看成一个人，那么情报预警系统、军事通信系统、指挥控制系统和火力控制系统就可以形象地比喻为"耳目"、"神经系统"、"大脑"和"手脚"。

1. 情报预警系统

情报预警系统是信息获取分系统，被称为"耳目"，主要是运用各种信息获取技术，进行情报收集、处理、传递和显示。主要设备有光学、电子、红外侦察器材和侦察飞机、侦察卫星以及雷达等。监视与侦察系统的作用是全面了解战区的地理环境、地形特点、气象情况，实时掌握敌友兵力部署、武器装备配置及其动向。

2. 军事通信系统

军事通信系统是信息传输分系统，被称为"神经系统"，主要是通过通信技术和网络技术实现信息的传输。通信系统通常包括由专用电子计算机控制的若干自动化交换中心，以及若干固定或机动的野战通信枢纽。手段包括有线通信（如海底电缆、光纤）以及无线通信（如长波、短波、微波、散射和卫星等），其功能是快速、安全、不间断地传输信息。

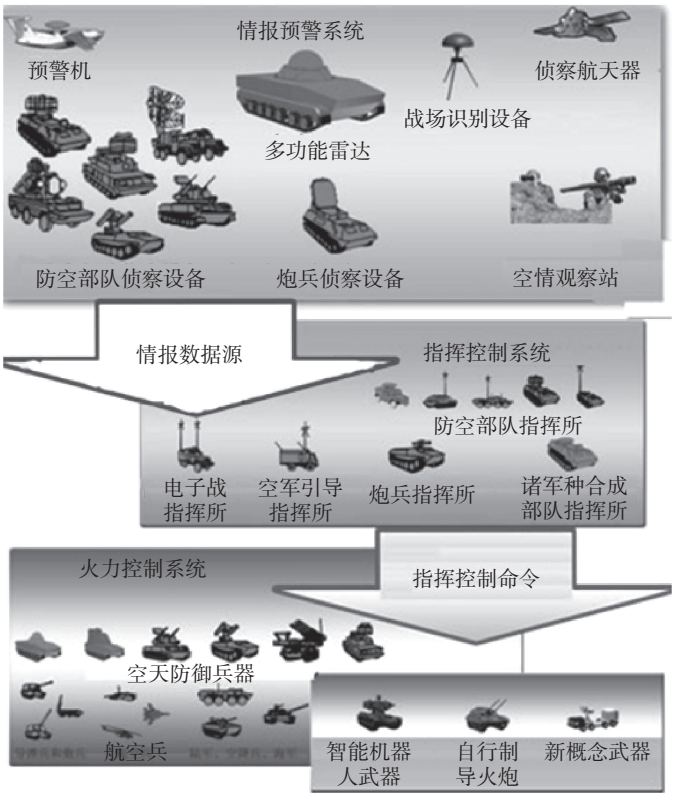


图8.1 指挥自动化系统体系结构

3. 指挥控制系统

指挥控制系统是指挥自动化控制系统的重要组成部分，被誉为“大脑”，负责军事信息处理、信息融合和决策支持等。指挥控制系统利用各级指挥机构中的计算机平台，对所搜集的信息进行自动化的处理分析，并科学地生成决策命令，以保障对部队的高效指挥。其任务是：(1) 接收信息并进行信息处理（格式转换、运算、综合、存储、输出等），为指挥人员提供形象、直观、清晰的战场态势信息（图像）；(2) 依据战场态势及有关作战规则、知识等，形成决策支持方案，进行模拟推演，为指挥人员决策提供参考；(3) 拟制作战计划，进行作战计算，分配作战任务，下达作战命令，及时准确地对部队或武器实施指挥控制；(4) 跟踪作战进程，适时调整作战计划和节奏；(5) 战况总结等。

4. 火力控制系统

火力控制系统被称为“手脚”，负责直接控制火炮、导弹等各种火力单元，向目标实施精准打击。

各组成部分各司其职，共同提高指挥自动化系统的作战效能，实现系统的综合集成。根据指挥自动化控制系统的特点，由于其任务、级别、军种、用途的不同，从而规模大小不一，功能各有特色，设备配置也不尽相同，但作战指挥控制过程基本一致，其系统作用过程如图8.2所示。从图中可以看出，指挥自动化系统是一个作战指挥人员在环中的、与系统外部作用对象（己方部队或武器）共同构成的闭环系统。

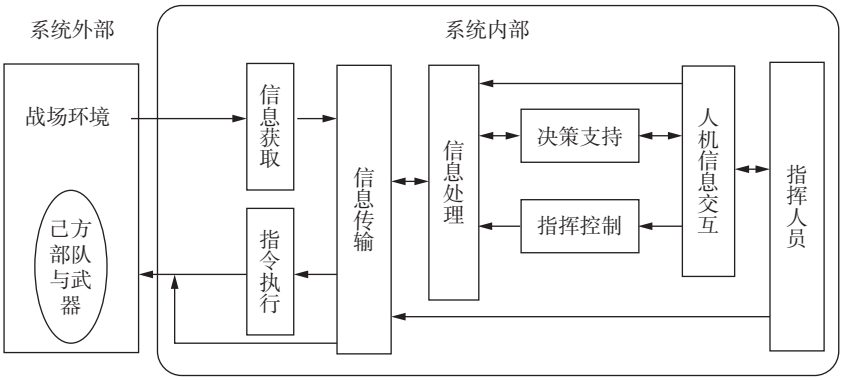


图8.2　指挥自动化系统作用过程

首先根据战场环境，利用各种信息获取手段，如空中侦察、卫星侦察和雷达侦察等获取情报，判断敌情。新情报信息经过通信设备传输到指挥所，并将其转化为数据存入计算机的动态数据库。然后，再由计算机对接收到的情报信息进行比较、分析、去噪、属性识别、威胁判断等一系列的信息处理后，给出对情报判断的结论，存储备查或分发到有关指挥人员的席位上，并在显示器上显示出来，供指挥决策时使用，同时上报上级指挥所和通报友邻部队及下属部队。

经过各子系统的配合，实现了军事信息处理的自动化，军事作战的一体化，战场打击的精准化，进而提高了整体作战的效能。

8.2　指挥控制系统

指挥控制系统在指挥自动化系统中起着核心作用。由情报监视侦察系统获取的情报，以及经由通信系统传递过来的信息，都在指挥控制系统中进行分析处理。对于这些信息，有的通过各种显示设备显示给指挥决策人员，有的由计算机进行加工计算，比对分析，去伪存真，生成表格和图例展示给决策人员，或通过一定的算法分析，帮助指挥决策人员判断作战情况，下达正确指令。

指挥控制系统的组成结构取决于军队的指挥体系。指挥体系是由各级指挥员及指挥机关组成的具有一定结构的网络体系，按军队指挥关系自上而下形成一个整体。指挥控制系统按军队指挥体系，同样自上而下形成一个整体，包括国家、战区、战役或战术军团、战术等层级，其中战术级指的是师以下级。美军自1995年开始建设全球指挥控制系统（GCCS），其结构如图8.3所示。

GCCS分为三层，最高层是国家汇接层，包括国家指挥当局、国家军事指挥中心和各战区总部及特种作战司令部等9个分系统。中间层是战区和区域汇接层，最底层是战区层，包括联合特遣部队、战区内各军兵种司令部系统。因此，指挥控制系统按指挥所级别分为战略级、战役级和战术级指挥控制系统。战略和战役级指挥系统还可按作战编成和业务需要，划分为若干“中心”，负责相关作战业务的计划、协调、组织、指挥。但是，作战指挥中心是这些“中心”的中心，担当着真正的作战指挥控制任务，其他各“中心”则担当作战保障和作战支持的任务。

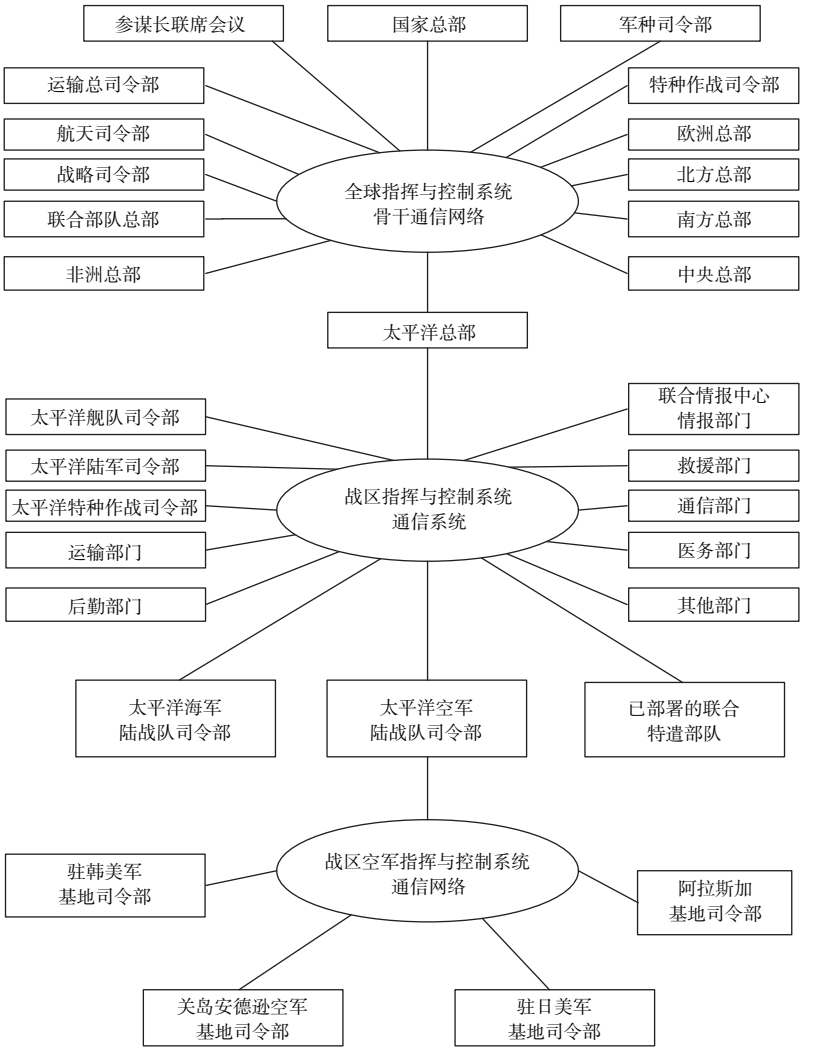


图8.3 美国GCCS体系结构图

8.2.1 美军战略指挥中心

美军有30多个主要的指挥中心，分布在世界各地，其中国家军事指挥中心、国家预备军事指挥中心、国家紧急空中指挥中心和国家舰载预备指挥中心是美国战略指挥自动化系统的“神经中枢”。

1. 国家军事指挥中心

国家军事指挥中心建于1962年，设在美国国防部五角大楼内。国家军事指挥中心供美国总统、国防部长和参谋长联席会议在平时和战时条件下指挥武装部队所用。指挥中心内存储有10多个战争总计划和60多个战斗行动方案，设有当前态势显示室、参谋长联席会议室、通信和技术室。参谋长联席会议通过这一指挥中心，用40秒的时间就可与国外任何一个或全部联合司令部进行联系或召开电话会议。

美国中央情报局、国家保密局、国务院、国防通信局以及联合侦察中心（见图8.4和图8.5）等有关部门和有关的办公室都派有代表在国家军事指挥中心工作。

图8.4　五角大楼卫星图

图8.5　五角大楼实景

2. 国家预备军事指挥中心

该中心设在马里兰州里奇堡的一个地下加固的设施内（见图8.6）。它与国家军事指挥中心相连，有较完善的情报收集、处理与显示设备，其功能大体上与国家军事指挥中心相似。它设有国家军事指挥中心的重要数据库，并且存放有进行常规战争和核战争的各种方案，可根据美军战备情况迅速增加人员，当美军进入二级战备后，它可立即承接全部军事指挥、控制任务。

该中心地下设施总面积达63 000 m^2，在300 m厚的花岗岩岩层下，四通八达的坑道建筑如同一个地下城，各种生活设施应有尽有，紧急时可容纳3000人以上。一旦全面核战争爆发，白宫和五角大楼被摧毁之后，这里将担负起美国国防部和总统栖身、指挥的双重功能，故有“备用五角大楼”之称。

这个中心可以说是美国绝密中的绝密。那些在上世纪50年代参加地下设施修建的人，也对自己的经历守口如瓶。因为他们当年是经过联邦调查局、国家安全局和国防部层层审查才有资格参加设施建设的，并且签署了终身不得向外透露任何情况的保密合同。“911”恐怖袭击事件发生后，该中心的设施全面启用，这也是该中心自建成以来的首次全面启用。

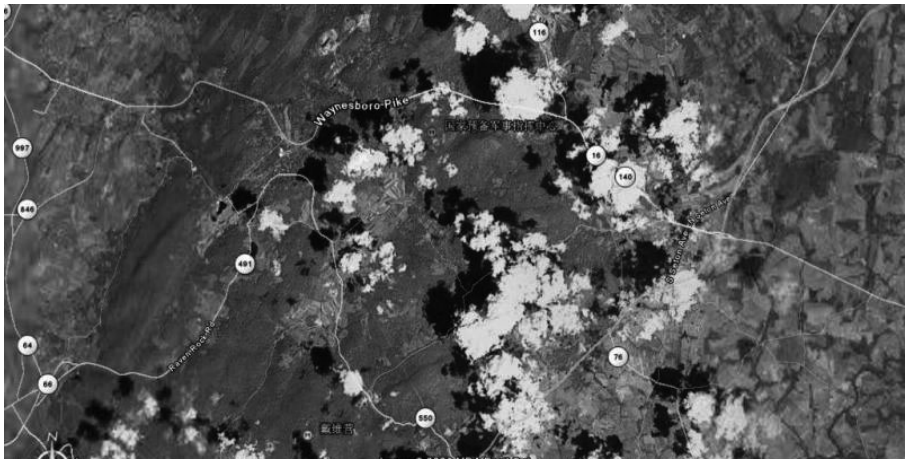


图8.6　国家预备指挥中心

3. 国家紧急空中指挥中心

国家紧急空中指挥中心（见图8.7和图8.8）也被称为机械指挥中心，运行着国家军事指挥系统，并运行着最低限度应急通信网的指挥机构。该指挥中心设在E-4飞机上，平时不参与指挥，只了解情况。当美军处于临战状态时，它升空待命。总统首次下达核攻击命令时，它取代陆地指挥中心行使对战略部队的指挥权。由于它在空中机动，是美军战略指挥自动化系统中生存能

力最强的一部分。20世纪70年代中期以前，国家紧急空中指挥中心由EC-135型飞机承担，70年代中期开始改用E-4A型飞机。使用E-4A型飞机的国家紧急空中指挥中心称为高级空中指挥中心。1985年全部改用E-4B型。这些飞机始终有一架在空中执勤，机上配有一位将军领班，另有一架处于战备状态，15分钟之内即可起飞。

图8.7 国家紧急空中指挥中心卫星图（1）

图8.8 国家紧急空中指挥中心卫星图（2）

E-4B在20世纪70年代后期由早期核战指挥机E-4A改装而来，机身选用波音-747客机，每架造价达2.58亿美元（见图8.9）。冷战时期，美军为了在核战中进行反击，准备在地面核战指挥所被摧毁后，迅速启用E-4B指挥美军进行核反击作战。美国空军目前有4架E-4B，配备第55联队，其中一架保持24h警戒状态，另一架随“空军一号”行动，以便于美国总统在全球任何地方能迅速登上该机。

图8.9 E-4B指挥机

4. 国家舰载预备指挥中心

该中心设在两艘战略指挥舰上，一艘是诺思安普顿号，一艘是赖特号。平时不参加指挥，只了解情况，当美军处于临战状态时，它出航待命。根据需要接替国家军事指挥中心行使对战略部队的指挥权。由于它在海上机动，所以它们和国家紧急空中指挥中心一样，具有较强的生存力。

5. 地下指挥中心

(1) 奥弗特空军基地

在奥弗特空军基地还有一个地下指挥中心，拥有美军最可靠、最有效、最先进的指挥控制和通信设备。其指挥控制和通信系统是美国的全球军事指挥控制系统中极为重要的组成部分之一，是美国总统指挥战争的一条重要指挥线。它设在内布拉斯加州奥弗特空军基地战略空军司令部总部地下（见图8.10），最下面的三层是一个由0.6 m厚的侧墙和1 m厚的顶盖构建成的三层长方体混凝土建筑物，平时有1000余人在里面工作，临战前夕和临战时完全封闭，但与外界的指挥通信不间断。该地下指挥中心的内部设施、防护设备完善，修备充足，可供800人工作和生活30天。

图8.10 奥弗特地下指挥中心卫星图

“911”恐怖袭击事件发生后，布什根据副总统切尼的请求没有立即返回华盛顿，而是乘坐“空军一号”总统专机迅速赶往奥弗特空军基地。在基地的地下指挥中心，布什可以指挥全球的美

军。该基地驻扎3架波音747-4B飞机，这种飞机经空中加油后可滞空72h。这样，即使情况再严重，也能确保国家指挥控制不间断和国家的稳定。

（2）韦瑟山绝密工程

韦瑟山绝密工程，又称“地下五角大楼”，该工程位于距华盛顿约90 km的弗吉尼亚州贝里维尔的韦瑟山中，20世纪50年代修建，耗资10亿美元，于1958年建成并投入使用。多年来，该工程经过不断扩建，以及抗震和防电磁脉冲加固技术改造，已成为先进而又坚固的美国政要“核战末日避难所”。该工程不仅是美国的国家预备军事指挥中心，也是国家紧急情况协调中心。该中心监视着世界各地所发生的重大灾难，如地震、核事故、核战争等。一旦受到核袭击，美国高级官员可以立即转移到这里，发布包括核报复和战后重建等各种命令。

韦瑟山工程占地约2000 000 m^2，地上约有30幢建筑物，设有微波中转系统、储水供水系统、污水处理系统、直升机场和指挥控制塔以及监控系统等。在防护层厚度75~90 m的坚硬岩石地下，修建了20幢钢筋混凝土建筑物，总面积54 000 m^2，紧急时可容纳数千人，里面有总统和政府要员的卧室，有地下水池和地下电厂，有医院、餐厅、娱乐和休息场所，有直通白宫的直拨电话，还有地下电车等。

“911”事件发生后，美国参、众两院议长以及其他法定的美国总统继承人，根据副总统切尼的命令，由美国海军陆战队特别保卫中队的直升机迅速疏散到该工程中并保护起来。这也是该工程自1965年11月以来的首次全面启用。

（3）夏延山地下指挥中心

夏延山地下指挥中心是美国全面核战争计划中最主要的“核战末日避难所”之一，同时它又是北美防空司令部和美国航天司令部的地下指挥中心。该中心位于科罗拉多州斯普林斯市西南郊600~700 m高的夏延山地下（见图8.11），1961年5月动工建设，1966年4月交付使用，该中心在山体深处的大洞穴中建有15座楼房。楼房的墙壁、地板、天棚、走廊和楼梯等结构全部由20~30 mm厚的钢板拼焊而成。每座楼房由几十根巨型弹簧支承，抗震性能良好。夏延山地下指挥中心是世界上可全天时监视全球航天器活动的唯一指挥控制中心，中心内部设施、防护设备完善，储备充足，能经受核武器的直接命中，紧急情况下，可供1800人工作和生活30天左右。

图8.11　夏延山地下指挥中心卫星图

“911”恐怖袭击事件发生当天，美国特工处曾建议布什总统躲进该中心，尽管布什最终没有躲进这里，但该中心却在这个事件后扮演了重要角色。在该中心中曾发布了“一号清空令”，要求所有美国上空和飞越美国上空的非军用飞机立即停飞。这道命令是20世纪60年代美国最高指挥当局为全面核大战爆发所准备的，这次是美国首次全面启动这道命令。

前面介绍的是美军的战略指挥中心。美军采用分级指挥的策略，除了部署战略指挥中心，还部署了战术指挥中心和战斗指挥中心，用于完成战役战斗的指挥工作。指挥中心的核心工作还是对获取的信息进行加工处理，显示给指挥官，为指挥官提供发布命令的态势参考和数据支持。下文结合美军的战术战斗指挥中心，主要介绍它们所担负的信息处理工作。

8.2.2　美军战术战斗指挥中心

战术战斗指挥中心担负战争中的核心任务，主要完成下列工作。

1. 信息展示

把情报监视侦察系统获取并经过通信系统传输来的各种信息，包括作战情报、敌我态势、作战方案、命令和命令执行情况等，自动进行信息融合与分析，提供当前战术环境信息，并报告敌友双方的行动和意图。最终用文字、符号、表格、图形、图像等多种形式，形象、直观、清晰地显示在各个屏幕上，供指挥和参谋人员研究使用。

信息展示基于信息可视化技术，可以实现地形地貌的二维标注及三维建模，可以实现敌我态势的动态实时展示，并可以对情报、监视、侦察系统获取的情报信息实现数字化、图形化、表格化的展示（见图8.12和图8.13）。

图8.12 指挥车内部战争态势显示

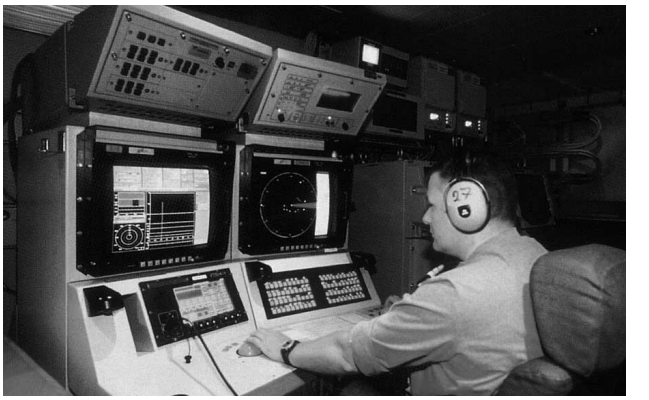


图8.13 舰船内部监视信息显示

2. 辅助决策

中心的计算机系统可以帮助指挥员选择方案，对威胁和目标进行分类，并按优先级排队，帮助指挥员了解情况，选择适当的对抗措施。通过计算机可以对各个方案进行逼真的推演，进行优劣对比，从而权衡各个方案的利与弊，从中选出最佳方案。

如图8.14展示了舰船指挥中心决策系统的工作情况，舰船有多坐标的不同范围的警戒观察显示屏。当可疑目标落入观测范围内时，系统自动发出警告提示警戒；当不明目标仍然前进，从而进入打击范围内时，系统会提示可实施打击。

图8.14 舰船自动判断提示实施打击

3. 发布指挥命令

实施指挥是指挥员的决策付诸实施的过程，是指挥周期中的最后一个环节。以电子计算机为核心的指挥自动化系统，可以使指挥员的决策及时准确地下达，并及时监督决策的执行情况。

8.3　情报预警分系统

情报预警分系统主要由情报侦察系统、预警探测系统和情报处理中心组成。它是指挥自动化系统中获取信息的单元。

虽然情报侦察和预警探测都是战场上重要的情报获取手段，但是两者在作战任务和使用目的方面又有所不同。预警探测可借助一系列传感和遥控探测手段，发现、定位和识别目标，并发出报警信号，而情报侦察则采用建立的各种情报网，通过观察、刺探、密取、窃听等手段获取情报。

8.3.1　情报侦察系统

情报侦察系统是一种典型的军事信息系统。情报侦察是指军事上为了弄清有关作战情况而使用秘密手段进行的活动。在信息化战争中，情报侦察系统是获取信息优势的前提和基础，为军事行动和作战指挥提供决策依据。

8.3.1.1　情报侦察系统的体系结构

情报侦察系统的体系结构如图8.15所示。其中，情报侦察系统由战略情报侦察系统、战役战术侦察系统、谍报人员情报侦察系统、平民情报搜集系统和电子战情报侦察系统等五大部分组成。

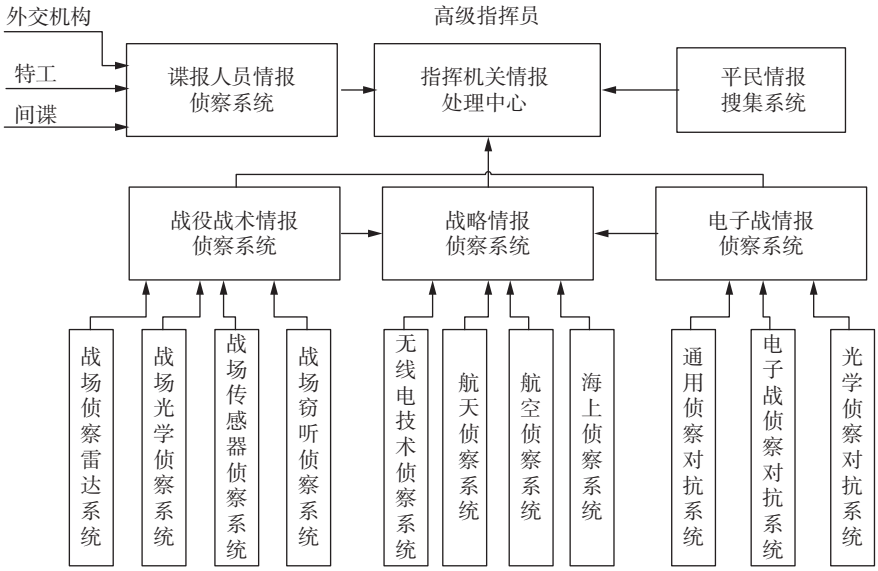


图8.15　情报侦察系统体系结构图

各种情报侦察系统获取的情报信息汇集到指挥机关情报处理中心后，都必须经过分析、整理或整编，将大量不完整、不精确、含糊和矛盾的信息通过计算机采用数据融合算法进行处理形成有价值的情报，这个过程也称为情报处理，再将情报传输、上报或分发给有关部门。其情报侦察处理功能如图8.16所示。

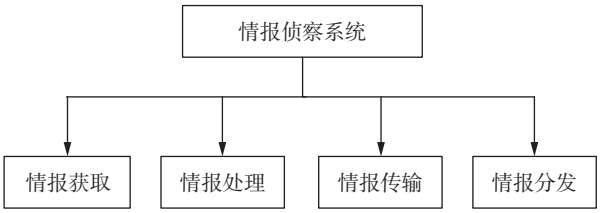


图8.16 情报侦察处理功能

谍报人员情报侦察和平民情报搜集虽是古老的情报搜集手段，但仍是十分重要的情报来源，因此在情报的获取方面仍发挥着重要的作用。平民情报搜集主要是指通过公开的多种媒体信息和社会舆论获取军事情报信息。谍报人员情报侦察是指向侦察对象内部秘密派遣或在侦察对象内部的秘密发展人员，以获取重要的情报。谍报人员情报侦察一般用于战略侦察，同时也广泛用于战役战术侦察。

战略情报侦察主要是获取国家安全和战争全局所需情报的侦察活动，同时它又是进行战略决策、制定战略计划、筹划和指导战争的重要保障，通常由高级指挥机关组织实施。所获取的情报主要包括：有关国家、集团和地区的战略指导思想及战略企图，武装力量的数量及其战略部署、备战措施、战争潜力、军政要人，以及社会、经济、外交、科技等情况，相关的国际环境及其变化对国家安全和战争进程的影响等重要情报。其重点是查明敌方战争直接准备程度，重点集团的集结地区，主要作战方向，核武器、生化武器的配置，以及作战开始时间、方式等影响当前战局发展最为急需的情报。同时战略侦察也为战役战术作战提供情报。

战役战术情报侦察是为获取战役战术作战所需的情报而进行的侦察活动，重点是进行战场态势的侦察，通常由战役指挥官或司令部组织侦察部队、分队或战斗部队实施，必要的时候可以发动平民参与情报的收集。主要工作包括对敌军的兵力部署、编制、装备、战斗编成、作战能力、作战特点、行动企图、指挥员性格、指挥机构、通信枢纽、军事基地、工事障碍、后勤和技术保障等进行侦察，查明敌方发起进攻或反击的时机、地点、规模和方向。

电子战情报侦察就是利用各种电子侦察系统，对敌方无意或有意辐射的电磁（或水声）信号进行搜索、截获、分析、识别、定位、记录并显示，并提取目标信号的各种技术参数和个体特征信息（包括：功能、类型、地理位置、用途以及相关武器和平台类别等情报信息），从中获得战略和战术情报。

8.3.1.2 典型系统

1. 天基侦察系统

美军通过向太空发射一系列的卫星来完成对地面的实时侦察和监控。这种侦察和监控从早期的光学成像方式发展到雷达方式和电子方式，各种侦察方式往往配合使用。

(1) 光学成像侦察卫星

美军最著名的光学成像侦察卫星为锁眼（KH）系列，其典型代表是KH-12。KH-12于1991年2月发射升空，已经发射多颗，曾经参加过近年来的几次局部战争。KH-12卫星分辨率高，运行在近/远地点398 km/896 km、倾角98° 的太阳同步轨道，卫星质量为17 000 kg，设计寿命为8年。它不仅可以在白天和黑夜发现敌方大部分裸露或伪装的地表目标，而且还能发现飞机发动机火焰和烟火等许多热源目标。此外，KH-12卫星还具有较强的机动变轨能力，可在需要的时候对热点和敏感地区进行抵近侦察。其缺点是在浓云层、雾、雨、雪等恶劣气象条件下，会受到严重的影响。

该卫星的星载有效负荷包括：大口径光学CCD照相机、红外热成像仪、监听微波通信信号和电话的信号接收机、数据中继转发器和天线等。

该卫星装有采用多光谱线阵器件和“凝视”成像技术的可见光CCD相机，具有多光谱成像和微光探测能力，目标分辨率可达0.1 m；为了提供优秀的夜间侦察能力，还装配了近红外/热红外成像仪，分辨率可达0.6 m。

（2）雷达成像侦察卫星

光学成像侦察卫星会受浓云层、雾、雨、雪等恶劣气象的影响，在这种情况下就需要使用雷达成像侦察卫星。美军最先进的雷达成像卫星是“长曲棍球”系列卫星。这种卫星的雷达波束能够穿透云层、烟雾、雨和雪，甚至对土壤、冰层、植被或其他掩蔽物也有一定的穿透能力。因此，无论白天和黑夜，也无论气象条件多么恶劣，“长曲棍球”侦察卫星都能够对地表实施全天候、全天时的侦察。

此类卫星的星载设备包括：SIR-D高分辨率合成孔径雷达等。

此类卫星的主要工作在微波波段，其波段为X、L双频段，“长曲棍球”成像分辨率可高达0.3 m，足以发现和识别诸如吉普车、坦克、舰船、导弹运输车等地面机动军事目标。甚至此类卫星还可发现藏于地下数米深处的军事目标，如飞机库、地下掩体等。

（3）电子侦察卫星

电子侦察是有别于光学成像和雷达侦察的另外一种侦察方式，它主要完成截收、侦听敌方电子信号的工作。“大酒瓶”系列卫星是美军第三代电子侦察卫星。

对于此系列卫星，每颗卫星均配有两部直径超过90 m的大型天线，可截收频率范围极宽的各种无线电通信信号和雷达信号。

此系列的卫星能够实现三种信号的截收、侦听，完成侦察工作：一是通过截收、侦听敌方无线电通信，可获取时效性很强的有关敌军部署等高价值军事与战略情报，并可测定其电台的信号特征及位置，为电子战奠定基础；二是通过截收敌方雷达信号，可测定其雷达的各种战术、技术参数及位置，为己方战机和导弹突防创造条件；三是通过截收敌方导弹试验的遥测信号，从中获取敌方导弹的性能，可大大提高己方的反导作战能力。

以上介绍的是美军在太空部署的各种侦察探测卫星。这些侦察探测卫星就像是天上的眼睛，帮助美军占据了情报监视侦察的制高点。

2. 空基侦察系统

美军除了在太空部署大量的侦察探测卫星，进行覆盖全球的军事侦察活动，还在空中使用各种侦察机展开监视侦察。其中，比较典型的侦察机有“U-2”、“RC-135”以及无人机等侦察机。

（1）U-2高空战略侦察机

U-2系列高空侦察机是美军进行照相侦察和探测地面目标电子信号的主要侦察机之一，绰号“黑寡妇”。U-2高空侦察机的最大平飞速度为930 km/h，最高可升到24 384 m高空，一次起飞能持续飞行8 h，侦察距离可达4700 km的范围。

该系列侦察机装备包括：全自动照相机8台，雷达信号接收机4部以及无线电通信侦收机、辐射源方位测向机和电磁辐射源磁带记录机等。在一次飞行中，U-2高空侦察机在15 000 m高空，清晰地拍下了宽为200 km、长为4300 km范围内的地面目标图像。

（2）EP-3E型“白羊座”电子侦察机

EP-3E型电子侦察机是美国海军的一种陆基信号情报侦察机，最高可飞行于8230 m高空，最大时速可达761 km/h，能连续飞行5400 km，一次起飞能持续12 h进行不间断侦察。

该侦察机配备了多种尖端的电子信息拦截系统，如信号收集系统、无线电方向探测器、自动频率测量接收机、多路无线电通信录音装置等电子侦察设备。它可以在740 km外的地方截获

到对方的雷达、无线电以及其他的电子通信信号，根据获取的信号判断对方的距离。该侦察机负责在公海海域执行搜集无线电和雷达频率等电子信号情报、探测跟踪舰艇的通信系统和侦察水下声呐阵地位置等重要军事任务。

（3）OH-58D“基奥瓦勇士”武装/侦察直升机

OH-58D侦察直升机主要用于野战炮兵观测以及为“地狱火”反坦克导弹和“铜斑蛇”制导炮弹攻击目标提供激光照射。其最大速度为222 km/h，航程可达556 km，可最高飞至6.25 km高空。其旋翼上方的瞄准设备包括高分辨率的电视摄像机、由120个成像单元组成的热像仪和用于激光测距的目标照射器。利用这些观瞄装置，OH-58D可以对目标坐标进行测算和测距，目标信息可通过卫星通信系统实时传输给地面炮兵，使之能及时发起精确攻击。

（4）美军新一代侦察机

美军的新一代侦察机主要有：“曙光女神”（Aurora）、“全球观察者”（Global Observer）、“广阔海域无人侦察机”（BAMS）等。“曙光女神”侦察机是高超音速战略侦察机，最快可达6倍音速，最高可到38.8 km高空，当代的任何战斗机和地空导弹都奈何不了它。“全球观察者”侦察机使用液氢引擎，可持续飞行一个星期，飞行高度达19.812 km，一次可侦察半径为960 km的区域，即面积为2 890 000 km^2，相当于1/3个中国；而“广阔海域无人侦察机”则使用顶部安置的360° 电子/光学红外全息录像机与电子引导雷达，一次可对半径为3000海里的区域进行全天候、全天时和全方位的侦察。

3. 地基侦察系统

地面侦察是战争中传统而又不可或缺的获取情报的基本手段，它不但可以弥补航天、航空侦察所提供的战术情报信息的不足，而且还能够验证航天、航空侦察所获情报信息的准确性，是及时了解敌方战役战术动向的有效手段。因此，美军除了在太空和空中部署监视侦察设施，在地面也部署了大量侦察装备，组成其地基侦察系统。这些地面侦察装备主要体现为侦察车、雷达、地面战场传感这三种不同的形式。其中，侦察车又分为装甲侦察车和无人侦察车。这里介绍装甲侦察车，无人侦察车将在8.6.1节介绍。

（1）装甲侦察车

地面装甲侦察车在弥补天基的侦察卫星和空基的侦察机盲点方面能起到重要作用，其主要侦察设备通常有大倍率光学潜望镜、昼用摄像机、热像仪、微光电视侦察系统、激光测距仪、小型战场侦察雷达、精确地面导航系统等。

美军现役装甲侦察车主要有M3履带式侦察车和M1127“斯瑞克”轮式装甲侦察车，其中履带式侦察车公路最大时速为68 km，最大行程为426 km，其配置的热成像瞄准镜不仅可进行全天候侦察，而且还可以自动追踪被锁定的对象。M1127“斯瑞克”轮式装甲侦察车的最大公路行驶速度为100 km/h，最大行程为502 km，其侦测系统包括红外光电探测仪、辐射和多用途射频传感器以及远距离化学污染监测设备等，能够在各种环境条件下对目标进行全天候、全天时、远距离的探测、识别和跟踪。

（2）战场监视雷达

战场监视雷达分为两种，其中一种是战场侦察雷达，另一种是火力侦校定位雷达。

战场侦察雷达是一种利用多普勒效应对野战环境下的人员、坦克、装甲车辆等地面活动目标和低空飞行器进行探测、识别和定位的雷达，最大作用距离从几km到50 km或更远，此类雷达的机动性强，能够全天时、全天候工作，而且还能发现在树丛里活动的目标。例如，美军的

现役雷达AN/PPS-15型号，该雷达能达到方位角覆盖范围360°、最大仰角34°，方位探测精度±12°，最远在1.5 km就可以探测到人类活动，在3 km就可以探测到车辆。

火力侦校定位雷达则是利用雷达波束去追踪敌火炮发射的炮弹，然后依据其弹道推算出敌火炮发射时的位置，为己方火力反击提供目标指示。火力侦校定位雷达也有很多类型。有用来侦察近距离目标的，例如，AN/TPQ-36火力发现者雷达对火炮和火箭炮的有效测程分别约为15 km和24 km；还有探测较远距离目标的，例如EQ-36增强型雷达，能够在混乱的环境里提供360° 的探测能力，对火炮的探测距离为32 km，对火箭炮为50 km。

（3）地面战场传感侦察系统

地面战场传感侦察系统是利用各种地面传感器，如震动、声响、红外、磁敏、视频、压力等传感器对地面目标进行检测、识别、分类、定位和跟踪的一种无源被动探测和监视设备。这种系统通过人工布设、飞机空投、火炮发射等方式，把大量的传感器撒布在敌方纵深地域内或敌方可能通过的地段和要道，对敌实施侦察。

美军现役地面战场传感侦察系统主要有远程监控战场传感器（RembassⅡ），即伦巴斯系统。其传感器有三种类型：震动/声响传感器、红外传感器、磁敏传感器。其中震动/声响传感器对履带式车辆的探测距离约为350 m，对轮式车辆约为250 m，对人员约为75 m；红外传感器对履带式车辆的探测距离约为50 m，对轮式车辆约为50 m，对人员约为20 m；磁敏传感器对履带式车辆的探测距离约为25 m，对轮式车辆约为15 m，对武装人员约为3 m。

为了应对日益复杂的地面战场环境，美军还装备了其他一系列地面战场传感侦察系统，如先进空中布设传感器系统（AADS）、未来战斗系统无人地面监视系统（FCS-UGS）、地形指挥官系统（OASIS2）、费尔康监测系统（FALCON）、模块化监视搜集观察单元（SCOUT）和“三叉戟”系统哨兵节点等，这些传感器可分别处理震动、声响、红外、磁场和视频等信号，而且通过与战术通信网络的对接，目标数据可实现远程甚至全球通信。

8.3.2　预警探测分系统

预警探测用于对目标的实时探测，探测信息用于实时指挥和控制。预警探测系统属全军公用部分，以实现各军兵种实时共享完整、精确、可信的军事情报。

8.3.2.1　预警探测的原理

预警探测系统的任务是实时探测、监视敌方各种目标的活动规律和动态情况，掌握敌对双方目标的分布态势，及时、准确地探测到任何威胁目标，迅速判断出目标的特性、种类等重要参数，并进行威胁度判断。

预警探测的基本原理是，利用多种媒介传感器，探测来自目标的电磁波、弹性波、应力等物理特征信息，通过数据融合技术，发现并监视目标。数据融合技术是指把各种来源的原始数据元素合并为一个单一的有意义的信息集合。结合军事角度，数据融合对来自多源的信息进行检测、关联、估计和综合等多级多方面的处理，以得到精确的状态和身份估计，完整、及时的态势和威胁估计。

8.3.2.2　典型系统

1. 天基预警探测系统

天基预警探测系统主要是指传感器平台位于卫星等天基运载平台上的预警系统，目前主要的天基预警探测系统是导弹预警卫星。

图8.17 导弹卫星示意图

美军现役弹道导弹预警卫星是第三代“国防支援计划”发射的卫星（见图8.17）。目前在轨5颗，卫星上的监测侦察设备主要有红外望远镜和高分辨率可见光电视摄像机等。其中红外望远镜有6000个红外探测单元，灵敏度很高，可及时发现导弹助推段尾焰的红外辐射。而电视摄像机则主要用于解决虚警问题，它能以每秒12帧的速度拍摄导弹尾焰运动轨迹的图像，地面站的工作人员可根据不同高度上尾焰的形状与亮度的差异判别出目标的真伪。国防支援计划弹道导弹预警卫星通常在导弹发射50~60 s内就能探测到导弹的情况，对洲际弹道导弹可提供25 min的预警时间，对潜射弹道导弹可提供15 min的预警时间。

美军正致力于发展“天基红外系统”（Space Based Infrared System，SBIRS）来替代国防支援计划。天基红外系统由高轨卫星和低轨卫星两类卫星组成，每类卫星都有数十颗。其中，高轨卫星主要用于导弹助推阶段的侦察与跟踪。每颗高轨卫星都配备紫外和可见光探测器，还配有用于大范围高速搜索的扫描型红外敏感探测器和用于小范围监视的凝视型红外敏感探测器，前者用于初始探测导弹助推段尾焰的红外辐射，然后将信息提供给后者进行精确跟踪。低轨卫星可在全球范围内跟踪导弹发射的全过程。每颗低轨卫星都主要配备宽视场高速捕获扫描型短波红外探测器、窄视场凝视型多谱段（中波、中长波和长波、红外及可见光）跟踪探测器这两种传感器。前者利用短波红外焦面阵列高速扫描南北半球，捕捉导弹在助推段喷出的尾焰；后者根据前者交接的信息，可对中段和再入段的导弹进行精确的连续跟踪和识别。据此，美军不仅可以精确计算出导弹的弹道轨迹，推断出其准确攻击地点，而且还可辨别出目标的真假，从而为美军的反导作战创造必要的条件。

因此，天基红外系统高轨卫星对导弹助推段的探测以及低轨卫星对导弹的中段和再入段的探测，大大加强了美军在全球范围内对导弹发射的全过程进行监视、跟踪和识别的能力。

2. 海洋目标监视卫星系统

海洋目标监视卫星系统是一种主要用于探测、监视海上舰船及潜艇活动，侦收舰载雷达信号和无线电通信信号的全天候侦察卫星。

在美军海洋监视卫星中，最著名的是“白云”系统，该系统已发展了三代，采用每组星座为1颗主星和3颗子星，一共4组星座包括16颗卫星（即4颗主星，12颗子星）的体制组网工作。目前，“白云”系列卫星已经逐渐被“联合天基广域监视系统”（SBWASS-Consolidated）计划替代。该计划由原“海军天基广域监视系统”（SBWASS-Navy）与原“空军/陆军天基广域监视系统”（SBWASS-Air Army）合并而成，其中前者是红外成像侦察卫星，其高灵敏度的CCD红外相机能够在全天候的情况下对水面舰船和水下潜艇进行侦察；而后者则由3颗卫星组成星座，其传感器主要为一部大型扫描雷达和一台电子侦察信号接收机，具有对飞机和水面舰艇的全天候侦察能力。

该系统的星载设备包括红外传感器、无线电侦收设备、侦测潜艇尾迹的微波辐射仪等设备。

该系统的主星主要包括成像侦察卫星，采用了可见光、红外、合成孔径雷达等多种成像传感器，以及用于探测核潜艇尾流的红外扫描仪和用于探测海况、确定海洋特性的微波辐射计等传感器。子星则为电子侦察卫星，通过3颗子星在空间排列成直角三角形，组网截收舰艇辐射的雷达信号，从而能够测定舰艇的位置、航向和航速等。

3. 空基预警探测系统

空基预警探测系统主要是指探测器放置于飞机、气球、飞艇等空中运载平台上的预警探测系统。

(1) RC-135高空电子侦察机

RC-135高空电子侦察机的巡航速度为860 km/h，最大航程可达12 000 km，飞行高度一般在15 km以上，续航时间超过12 h，如果借助空中加油，则巡航时间可达20 h。

RC-135高空电子侦察机有多种型号，其中RC-135 S型号的主要任务是搜集弹道导弹情报。RC-135 S配备有可见光、中频红外传感器和长程激光测距系统等高精度光电监视系统和机载电子侦察设备，可以跟踪导弹的飞行状态，收集、处理和分析导弹制导的电波频率及相关信息，能在418 km外判定导弹的发射点、导弹发动机的熄火点，并精确计算出导弹轨迹和着陆点；还能依据导弹弹头与大气层摩擦时产生的闪光，判断出弹头的材料、速度以及是否具有机动能力。

RC-135 V/W型号的重点任务是实时侦测空中的各种电磁波信号，相关技术：配备有多种传感器，可以测量世界上各种雷达的频率等技术参数，并对雷达进行定位，方位可精确到 ± 1°；可以广泛侦收语音、电传、电报等无线电通信信号，在10 km高度可侦测到60~800 km距离的电台并确定其位置。同时，其机载红外探测器和前视雷达可探测238~370 km范围内的目标，并可分辨出3.7 m长的物体。

(2) E-8C雷达侦察机

E-8C雷达侦察机是美军一种主要用于搜索、监视地面目标，并可对地面战场实施有效管理的重要空中侦察指挥平台。其最大平飞速度为1010 km/h，可最高升到12.6 km高空，续航时间为11 h，最大航程为6920 km。

这种侦察机配备有能用于联合监视并攻击目标的APY-3相控阵雷达系统。该雷达能够在任何气象条件下对地面目标进行探测、定位和跟踪。其中，合成孔径雷达用于对地面固定目标的监视，该雷达波束每35 s对160 km × 180 km的区域进行一次扫描，可获得高分辨率的敌方阵地和地面固定目标图像；而脉冲多普勒雷达则主要用于对地面机动目标或直升机等低空慢速运动目标的探测、定位和识别，对运动中的坦克和装甲车的识别距离可达150 km，而且该雷达还能区分出轮式和履带式车辆或其他地面机动目标，并且能判断出其运动方向和速度，从而可以了解敌方作战行动的意图。

4. 陆基预警探测系统

随着技术的不断更新和发展，陆基预警探测系统所包含的内容越来越广泛，手段也越来越多样。陆基预警探测系统主要由各种地面固定或机动式雷达、电子侦察设备、光电探测装备等组成，主要包括无线电通信侦察和雷达侦察等。

“铺路爪”AN/FPS-115潜射弹道导弹预警系统是应对洲际导弹威胁而研制的远程预警系统，主要用途是担负战略性防卫任务。该雷达采用双面阵天线，工作频率420~450 MHz，能对4000 km以内的低弹道潜射导弹和5500 km以内的高弹道潜射导弹做出准确、及时、可靠和有效的预警，能及时提供导弹发射点、弹着点、位置和速度等早期预警数据，并具有同时跟踪、检测和识别多目标的能力（见图8.18）。

图8.18　“铺路爪”AN/FPS-115

8.4 军事通信系统

军事通信系统是指挥自动化系统的“神经网络”，也是国家和军队的重要基础信息设施。它主要用于将获取的军事信息传递出去，并负责装备与装备之间、装备与指挥所之间、指挥所与指挥所之间的军事信息交互与分发。美军的通信系统可分为战略通信系统和战术通信系统。

8.4.1 美军战略通信系统

战略通信系统把战略预警探测系统和战略指挥中心联接起来，并在各指挥中心之间传递信息。战略通信系统担负着非战时的军队应急通信以及战时决策级指挥通信联接服务。

美军战略通信系统包括通用和专用两部分。通用通信系统包括国防通信系统、国防卫星通信系统和最低限度紧急通信网；国防通信系统由自动电话网、自动密话网和国防数据库组成，线路总长达67 200 000 km，能把世界上100多个地区的3000多个指挥所连接起来，主要用于保障美国总统与国防部长、参谋长联席会议主席、情报机关和战略部队的通信联系，也可以为战术通信提供通信枢纽；国防卫星通信系统DSCSⅢ是美国战略、战术共用的卫星通信系统，它由14颗卫星和70多个地面站组成，其中14颗卫星为地球同步卫星，位于赤道上空，主要工作在超高频波段（最晚发射的4颗卫星上增设了特高频通信设备），可为位于东太平洋、西大西洋、东大西洋、印度洋和西太平洋等五个区域的美国陆、海、空三军提供加密且可靠的全球通信服务。该系统（见图8.19）承担美国战略通信70%的通信量，用于传递战略指挥信息情报数据、高度优先的战略预警信息和特种信息等，是美国全球指挥控制系统远程战略通信的支柱；最低限度紧急通信网是专供国家最高军事指挥当局在核条件下把美国核战争计划的命令传送给美国在全球的核部队，并接受核部队回报执行命令的情况，主要采用甚低频到特高频的所有通信手段，以保障通信的可靠性和生存能力；专用通信系统主要包括：空军卫星通信系统，极低频对潜通信系统，机载甚低频中继机通信系统，战略空军司令部通信系统等。

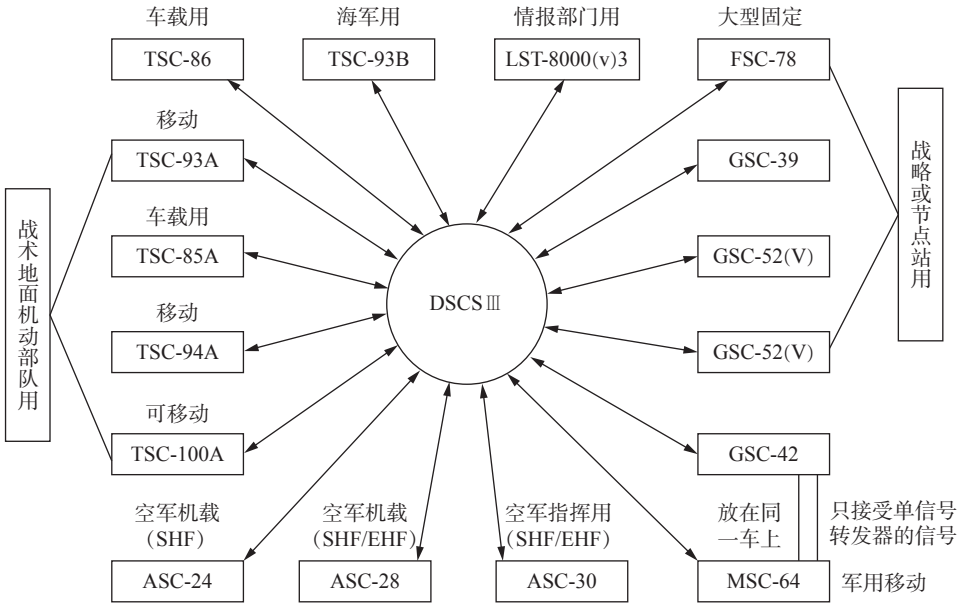


图8.19 DSCSⅢ卫星通信系统及主要地面站型号

8.4.2　美军战术通信系统

战术作战通常指具体地区内的战斗。战役战术通信系统通常是为了保障师以上部队遂行战役作战和师以下部队实施战术行动。战役战术通信系统网中的固定通信设施是战略通信网的组成部分，而机动部分则是战时在战区开设的通信设施，以野战通信装备为主，可以根据需要随时接入战略通信系统。战役战术通信系统主要包括地域机动通信系统、移动通信系统、战术互联网、战场数据分发系统、战术卫星通信系统以及数据链等。其中，数据链作为一种重要的军事战术通信系统备受瞩目。

对于数据链，有很多不同的称谓。为了满足战术信息传输的需要，美国等北约国家于20世纪50年代开始研制和装备“战术数字信息链路”（Tactical Digital Information Links，TADIL），简称数据链。数据链在系统各个用户之间依据共同的通信协议和信息标准，使用自动化的无线（或有线）收发设备传送和交换战术信息。在数据链中，包含分发信道设备、通信协议和消息格式标准，数据链是一种军事上特有的通信应用。

通俗地讲，数据链就是互通数据的链路，而在军事上所说的数据链就是一张数据网。正如互联网，只要有一个数据终端就能从这个数据链获得自己所需的信息，同样也能用终端在这个数据链路网里添加东西。例如，一架无人机发现某个地点有一支部队，首先它可以通过数据链来识别该地区是否有自己的部队，进而识别该部队是敌人还是友军，而当识别为敌人之后就能通过数据链发布消息。这样无论是地面陆军坦克、空军战机或后方指挥部，都会获得这一消息。而这个网络功能可以无限细化，比如组建一个专门的救护系统，如果战场上有人受伤就能立刻把伤员的位置、受伤部位、伤员信息等都发布上去，而专门的救护部门则不间断地检测是否出现新的伤病信息。只要出现伤员信息，就能根据报告的伤员相关信息准备好相应的急救装备，然后联合敌我双方的分布数据来选择一条安全的通道奔赴伤员所处位置，进行快速有效的救护。而且，甚至可以让已经出发在空中的救援直升机顺路接收新的伤员，这对于提高救护效率是一个巨大的帮助。在美军近几场局部战争尤其是伊拉克战争中，数据链发挥了重要作用。数据链已成为三军联合作战中进行实时或近实时指挥控制、战场态势信息分发的主要手段。

战术数据链的建设始于20世纪50年代，首先是装备于地面防空系统和海军舰艇，之后才逐步应用到飞机上。到目前为止，已有多种战术数据链问世，大致可分为以下几类：一类是态势感知数据链，也是一种较为通用的数据链，用于各军兵种多种平台之间交换不同类型的最新信息、满足多样化任务需求，一般工作在低频，波长较长，数据率较低，主要是传输格式化报文信息，包括Link1、Link4A、Link11、Link16和Link22等。一类是情报、监视和侦察（ISR）数据链，用于传输各种图像情报和信号情报信息，一般工作在高频，波长较短，数据率较高，能实现视频和高分辨率图像的高速传输，包括通用数据链（CDL）和战术通用数据链（TCDL）等。此外还有专门为新型设备研发的数据链，如无人机数据链、弹药数据链等类型。

8.4.2.1　美军态势感知数据链

美军将数据链称为TADIL，Link是北约对数据链的称呼。目前通用数据链以Link系列为主，应用的Link系列数据链主要有Link1、Link4A、Link11和Link16等，其功能特点如表8.1所示。

表8.1 各种数据链技术的功能特点

类型	组网	范围	工作频率	设备
Link1	点到点	陆上通信线路	—	数据缓冲装置实现了防空数据的自动交换
Link4A	点到多点；轮询技术（点名呼叫）	仅供所有航空母舰上的舰载飞机使用	特高频	特高频无线电台、调制解调器、密码设备、数据处理器和用户接口设备
Link11	多点到多点	支持战斗群各分队之间的交换态势信息的海军战术数据系统	高频和特高频无线电	战术数据系统、保密设备、数据终端、高频和特高频电台
Link16	立体网络	联合战术信息分配系统	具有抗干扰能力的特高频无线电	综合通信、导航和敌我识别，保密性能好

Link22由北约国家共同开发，是用以取代Link11的新一代数据链系统，也称北约改进型Link11。它是一种保密、抗干扰的超视距战术通信系统，主要应用于海上舰队，可在陆地、水上、水下、空中或太空各平台之间，进行电子战数据交换，以及指挥控制指令与情报信息传递。为了在信息格式上与Link16兼容，Link22采用了由Link 16衍生的信息标准以及Link16的结构和协议。Link22与Link16一样也采用TDMA技术，在高频和超高频频段采用跳频模式以提高抗干扰能力。Link22的传输速率比Link16慢一些，但它通过自动化网络管理技术提供动态配置网络的能力，同时还具有电子战能力。Link22系统在高频和特高频频段分别提供300海里和200海里的通信覆盖范围，使用中继设备以后，可分别达到1000海里和300海里。

以上介绍的是美军的通用型数据链，用于各军兵种多种平台之间交换不同类型信息、满足多样化任务需求。除了通用数据链，还有一些新型的专用数据链，用于完成一些特定的通信任务，下面介绍美军的一些新型数据链。

8.4.2.2 美军新型数据链

随着网络中心战概念的实施，现有数据链已无法满足远距离、高动态、大容量、低时延等信息传输的要求。为此，美军正在研制各种新型数据链，如情报监视侦察（ISR）数据链、网络数据链、弹药数据链，以及激光通信数据链等。

1. ISR数据链

虽然Link数据链能够完成传输信息的任务，但随着信息技术的发展，传送的信息量越来越大，Link数据传输速率无法满足ISR系统中的图像实时高保真传输的要求。因此，美国国防部于20世纪80年代开发出通用数据链CDL，并于1991年命名为ISR标准数据链。CDL的最大特点是宽频带和通用性，标准数据率可达10.71~274 Mbps，并正在向548 Mbps迈进，传输速度更快。CDL是一种全双工、抗干扰的微波通信系统，是一系列可以互操作的、可供各种特殊应用平台选择使用的数据链。通过CDL平台可将光电、红外、合成孔径雷达等传感器所获取的图像和视频信号等信息以视距或经由中继超视距传输到地面控制站或舰艇。目前CDL主要用于侦察机、无人机等空中平台，已装备于U2侦察机、战术飞机侦察吊舱等各种主要的ISR平台中。

为了将多种配备CDL的终端都能连接上网，以更好地进行网络控制，美国空军实施了多平台通用数据链MP-CDL计划。MP-CDL能在网络环境下为不同装备之间提供经济有效的视距、宽带、空对空与空对地的数据链路。它具有许多先进通信能力，包括自主修复网络、IP路径选择、自适应传输功率、自动信号获取和抗干扰等。MP-CDL机载终端能够同时进行视频电话会

议、传输高清晰度视频、通过互联网协议传送话音到公共交换电话网，以及访问互联网和收发电子邮件等任务。作为一种网络中心数据链路，MP-CDL还具有向地面作战人员提供实时运动视频的能力。MP-CDL工作在Ku波段，将来可能扩展为X波段和Ka波段，理论传输速率为10~274 Mbps，具有很强的抗干扰能力。它采用“网络广播”和“点对点”工作模式。当采用网络广播方式工作时，可以同时向32个（最多50个）用户发送信息；采用点对点工作模式时，可在2个平台之间进行高速数据交换。MP-CDL将在机载与地面ISR平台之间提供网络中心数据链路，成为军方装备的第一个完全网络化的CDL。在改进的“联合星”和新的E-10飞机上都将装备该数据链，其地面站将作为全球信息栅格的网关。

前面介绍的CDL系列数据链主要适用于战略装备，不适用于战术平台。随着无人机在美军的广泛应用，美国国防部又开发了成本更低、体积更小的战术通用数据链TCDL。

战术通用数据链TCDL，又称“鹰链”。TCDL与CDL兼容，相互之间可以实现近实时链接与互操作。TCDL是美国海军构筑网络中心环境下先进ISR网络的重要装备，目前，P-3C巡逻机、RC-12侦察机以及“猎人”战术无人机都装有TCDL，“先锋”、“捕食者”、“影子-200”战术无人机及MH-60R直升机也将装备TCDL，其外观如图8.20所示。

图8.20　TCDL终端

2. 小型无人机数据链

前面我们介绍了ISR数据链，TCDL已经用于战术无人机。与TCDL不同，小型无人机数据链主要用于手持发射无人机。手持发射无人机比战术无人机级别更低，其终端通常使用背负式地面站，代表产品为MRS-2000背负接收系统。MRS-2000为远程接收站的替代品，是一种全数字化、坚固耐用的便携式智能计算机，用于搜集和分发来自无人机的视频图像和遥感数据。MRS-2000工作波段包括C波段、L/S波段和Ku波段，有效距离从8 km（使用全向天线）到48 km（使用定向天线）。美军“先锋”无人机已配备MRS-2000，并在伊拉克战争中成功使用。

3. 弹药数据链

为了与武器平台之间沟通瞄准信息，精确制导弹药也需要装备数据链。增程数据链（ERDL）是较早一代的弹药数据链，可以传输从导弹发射到击中目标期间的视频图像，是为AGM-62“白眼星”导弹研制的。现代弹药数据链的功能更多，比如可以使飞机能够控制飞行中的弹药并重新瞄准。现代弹药数据链除了具有飞行中重新瞄准能力，还具有情报监视侦察功能。

洛·马公司正在为“未来作战系统”研制非直瞄发射系统（Non Line Of Sight-Launch System，NLOS-LS），包括精确攻击导弹（PAM）和待机攻击导弹（LAM），其中就包括其弹药数据链。LAM可以通过数据链传输目标情报，其数据链在1.2 MHz时的传输速率为900 Kbps~2.4 Mbps，在4 MHz时的传输速率为8 Mbps。

4. 网络数据链

网络数据链（NDL）的工作频率很高，其频率在2 GHz以上，为获取的情报信息传输提供了高速的通信链路。网络数据链主要用于地面环境，将各种地面平台尤其是“未来作战系

统”联成网络。网络数据链的核心技术包括定向网络波形（DNW）技术和战术瞄准网络技术（TTNT）。

定向网络波形技术是波音公司于2004年公布的战区移动定向通信技术。它采用时分多址技术，工作频率15 GHz，每条链路的数据速率为200 Mbps，通信距离为250海里，可同时与另外6个节点通信。与其他下一代波形技术一样，它也使用IPv6和移动自组网或网状网络，避免了使用分散网络带来的平台问题。DNW可使战区中的士兵从同一来源（如指挥中心、战斗机及无人机）接收所需的各类信息。

战术瞄准网络技术是下一代数据链的代表技术，它兼容了Link16数据链，最大特点是有效扩充了网络容量，并提高了用户自组加入网络的速度。使用战术瞄准网络技术可以组织包含2000个成员的大型网络，其网络容量高达10 Mbps，信息延迟为1.7 ms，网络管理协议更新速率和新用户进入移动自组织网络的时间均为3 s，作用距离为121海里，每个用户可以使用的通信容量为2.25 Mbps。

5. 激光通信数据链

激光通信的容量高达Gbps量级以上，是真正的极高频传输。例如，美国“转型卫星通信”系统（TSAT）将在卫星之间以及卫星与高空载人飞机/无人机之间提供高达10 Gbps的激光链路。为了利用TSAT传输的信息，美国空军正在实施“多通路激光空间终端”（MALST）计划，由诺斯罗普格鲁曼公司开发的终端技术将建立基于固体激光通信的10 Gbps星际链路。同时，美军还在开发机载激光通信终端技术。该终端可接收在轨卫星发出的信息，然后将其传输至战术级别的机载设备，以比目前射频系统更高的速率向用户提供ISR数据。

现今，数据链管理已日益提上日程。随着数据链应用的增加，各种链路之间的互联互通以及资源信息共享的最大化，都要求必须有一种数据链管理方案，能够对数据链和网络服务进行自动化智能控制，以提高作战效率。例如，美军委托通信公司开发了网络平台通信管理器（IPCM）方案。这是一种多数据链管理方案，在2004年8月美军举行的远征部队试验演习期间做了演示。IPCM可以实现对数据链和网络服务的自动化智能控制，支持多平台通用数据链、多任务战术通用数据链与航空通用传感器，以及未来的网络中心协同瞄准计划、多任务海上飞机及战术机载侦察系统。利用IPCM技术，可以连续监视与地面站关联的飞机位置，预测保持TCDL链接所需的数据传输速率及平台天线的变化，可以大大降低数据链中断率，明显改善无线网络的性能，显著提高数据链的可用性。

无论战略通信系统还是战术通信系统，都需要依靠种类丰富的信息传输技术，如光纤、短波、微波、卫星通信等，从而在各种环境中保持高速、可靠的通信。同时，在多个指挥所、情报监视侦察设备、火力控制设备之间部署各类有线或无线网络，通过网络实现信息的传递。在网络中传递的数据，需要有一定的格式控制及协议控制，军用数据链完美地解决了信息传输中的控制工作，实现了装备与装备之间、装备与哨所之间、装备与指挥所之间的有效通信。

8.5 火力控制系统

火力控制系统（Fire Control System，FCS）又简称为火控系统，泛指控制火炮、导弹、鱼雷等武器瞄准和发射的成套设备，是武器系统的重要组成部分，最终目的是快速、精确、有效地击毁敌方目标，并保存自己。

8.5.1　火力控制系统的功能和组成

8.5.1.1　火力控制系统的功能

火力控制系统的主要功能是控制武器设备对目标进行有效攻击。实际的火力控制系统可能各不相同，但其功能基本相同，归纳起来有以下几点。

(1) 接收目标指示信息和载体的参数信息，对目标进行定位跟踪；

(2) 预测武器的战斗部或弹丸与目标的相遇点，解算命中目标所需的射击（引导）诸元；

(3) 完成发射瞄准和适时开火，控制发射的全过程。

8.5.1.2　火力控制系统的组成

火力控制系统有多种类别，广泛应用于陆、海、空以及第二炮兵各军种的多种兵器中。归纳起来，每一种火力控制系统都可划分为5个子系统（见图8.21）。

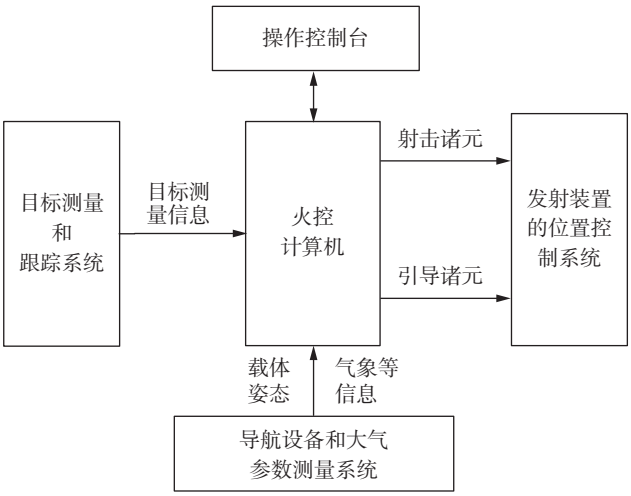


图8.21　火力控制系统框图

1. 目标测量和跟踪系统

该系统包括测量和跟踪目标的设备，其任务是测量目标的距离、方位、高低角（俯仰角）或其各阶变化率，目标的速度、航向、距变率（径向速度）和横移率（即方侠线垂直方向上的移动速度），并将这些数据送至火力控制计算机。常见的测量跟踪装置有光电观瞄设备、被动式红外跟踪装置、雷达、激光雷达等。

2. 导航设备和大气测量系统

导航设备实时测量武器载体的姿态参数和运动参数，大气测量系统测量风速、风向、高差、气压等参数，并将这些特征参数传送至火控计算机进行参数计算。

3. 火力控制计算机系统

火控计算机是火力控制系统的核心，其主要任务是接收测量装置和跟踪装置所测得的目标数据（如敌我距离、方位、俯仰角等），接收导航设备和大气测量系统所测得的武器载体的姿态参数、运动参数及大气参数，计算目标移动速度、方向、位置、加速度，以及武器射击诸元，如导弹自控时间、武器的发射架瞄准角等。

4. 发射装置的位置控制系统

该系统的任务就是接收火控计算机计算的射击诸元，定位发射装置或直接给武器装定某些射击诸元。

5. 操作控制台

操作员通过控制台按钮、键盘来操作火控计算机，完成相应的计算和控制动作。操作控制台还可通过数码管、指示灯或显示器，把文字、图像、声音等交互信息以多媒体手段直观形象

地提供给操作员。操作员可通过控制台控制武器平台发射，还可以实现显示设备自控状态、指示故障部位、指导模拟训练等功能。

8.5.2 火力打击网络

火力打击网络由各种不同的武器单元相互联接而成。火力打击网络不仅将武器系统联网，而且能发挥整体合力。信息与火力的有效融合，使联合火力打击成为主要手段和作战样式，同时也是达成战争目的的关键因素。因此，在指挥控制与火力控制一体化系统中，各种空基、陆基和海基武器系统和用于指挥控制这些武器系统的各种软硬件，构成了一个火力打击网络，它的任务是有效地利用战场感知信息达成预定的作战效果。

火力打击网络在指挥控制与火力控制一体化系统的作用下，通过对目标的杀伤、破坏，实现指挥决策目标，同时保护己方的各种资源。

火力打击网络的关键技术主要包括：火力打击网络的信息化技术、自动化技术、武器共架发射技术、综合火力控制技术、火控系统组网技术等，本节将介绍火力打击网络的信息化技术和自动化技术。

1. 信息化技术

为了在武器平台实现信息与火力的一体化，首先要发展信息化精确制导弹药和智能化弹药。精确制导弹药能在敌人火力网之外发射，自主识别和攻击目标。智能化弹药能在各种条件下利用声波、无线电波、可见光、红外、激光甚至气味或气体等一切可利用的目标信息，自主选择应攻击的目标和攻击方式。其次是使作战平台（例如装甲车、火炮、导弹发射装置、直升机等）信息化，在这些装备上安装大量信息系统，如车际信息系统、自动化火控系统、指挥员综合显示系统、导航与目标瞄准系统等，使作战平台具有与传感器、指挥中心和友邻联网的能力，做到信息共享、迅速反应、精确打击。最后是发展数字化士兵装备，使士兵由过去单一的战斗员转变为集侦察、通信和作战三位于一体的多能士兵。

信息化武器系统的组织、指挥控制与火力运用，必须在信息的主导下才能发挥最大的作战效能。信息化能提高火力打击精度和打击速度，摧毁敌方火力点，减小威胁，增加己方生存力，保持火力，实现持续战斗力。若用火力摧毁敌信息系统，则敌方信息力降低又抑制了敌方武器系统作战效能的发挥，从而更有利于己方武器系统作战效能的发挥。

火力打击网络的信息化就是要在武器平台实现信息与火力的一体化。信息与火力高度融合的武器系统，能够使预警侦察、指挥控制、精确打击、毁伤评估、战场管理等领域的信息处理网络化、自动化、实时化。

2. 自动化技术

在实现了信息的采集之后，指挥控制中心就对各种信息实现自动化的信息融合、威胁判断、攻击排序和目标分配，进而将这些信息通过战术数据链自动传递给各个火力平台。这显然就要求火力平台在接到指挥系统发送的目标数据和命令信息后，能够在无须人工干预的情况下，将这些信息自动传输给火力控制系统，控制武器对来袭目标进行自动打击，从而实现从发现目标到打击目标的全过程自动化。

8.5.3　火力控制新技术

8.5.3.1　多目标攻击火力控制技术

多目标攻击，就是一架飞机同时攻击空中多个目标或两架以上飞机协同攻击空中多个目标。多目标攻击需要对空中多个目标同时进行跟踪、识别、火力控制计算和对多枚导弹同时进行制导。在多机协同进行多目标攻击时，还要对攻击空域和目标进行合理分配，分配的结果应当使整体杀伤概率最大并尽可能避免重复攻击。

1. 多目标火力控制功能

多目标火力控制功能包括：对雷达提供的多目标信息进行坐标转换、平滑、滤波、外推、预测；在此基础上进行目标威胁判断，攻击排序，火力分配，导弹允许发射区计算、飞机操纵指令计算、雷达照射兼容性检查和扫描中心计算，攻击逻辑计算以及导弹发射条件和飞行任务计算；导弹发射后，计算无线电修正指令并传送给雷达，分析空情态势，确定是否改变攻击计划；判断目标是否被摧毁，确定是否进行再攻击或退出攻击、返航。

2. 多目标火力控制的关键技术

(1) 目标威胁判断、攻击排序

目标的威胁程度与目标的类型、方位、距离、高度、速度、进入方向、战略意图，以及所携带武器的种类、数量、精度、破坏威力等因素有关。威胁判断的任务，就是根据上述因素中可获得的目标情报数据，按照某种规则计算目标的威胁程度，并对其进行排序。

(2) 载机的最佳导引航迹的寻找和操作指令计算

多目标攻击载机的最佳导引航迹应满足以下条件：

- 保证飞行员指定的目标受到攻击；
- 保证最危险的目标受到攻击；
- 保证选定的攻击目标周围具有公共发射区，并受到攻击；
- 保证能攻击到最大数量的较危险目标；
- 保证雷达照射和无线电修正指令照射兼容性；
- 保证按攻击排序的顺序攻击目标；
- 发射的导弹到达攻击排序中第1个目标的时间最短。

(3) 雷达照射和无线电修正指令照射兼容性检查程序

所谓照射兼容，就是要确保受攻击的目标，在导弹拦截（或命中）目标之前，一直可以受到雷达的照射。

(4) 多枚导弹同时制导

在多枚导弹发射后，在最后一枚导弹截获（或命中）目标前，火控计算机需要控制雷达跟踪被攻击的（和未被攻击的）多个目标，并向多枚导弹发送无线电修正指令。跟踪目标和指令发射的时序分配（保证每枚导弹都能接收到无线电修正指令信号）、指令保密以及抗干扰，都是同时制导多枚导弹成功攻击目标的关键。

(5) 多目标攻击显示技术

多目标攻击时，综合显示系统应将所跟踪和攻击的目标数据分别显示在战术信息显示器上，同时显示器还应显示目标的威胁信息、航向信息、载机操纵信息和飞行状态信息。

3. 多目标火力打击典型应用

一直以来，各国战斗机的发展需求都以空对空作战为主，尤其是美国的第四代战斗机F-22（空中优势战斗机），其“先敌发现、先敌发射、先敌摧毁”的超视距多目标攻击能力主要也是体现在空对空的作战上。下面以F-22所具有的特点来分析现代战机火力控制系统的情况。

（1）超视距多目标攻击

早在20世纪70年代，美军就已开始研究超视距多目标攻击技术，并已在第三代战斗机上应用。目前第三代战斗机，如F-15、Mig-31M 和Cy-35等已具备超视距多目标作战能力。第四代战斗机进一步发展了超视距多目标攻击技术，普遍采用主动雷达制导技术，发展了更为先进的第四代超视距多目标空空导弹。

例如，美国的远距空空导弹“不死鸟”系列的AIM-54C，最大射程为150 km，可同时攻击6个目标。俄罗斯的远距空空导弹P-37，可同时攻击6个目标，最大射程为200 km。据报道，俄罗斯正在研制的超远距空空导弹AAM-L，最大射程可达400 km。超视距多目标攻击将成为第四代战斗机的主要空战方式。

（2）多机协同多目标攻击

多机协同多目标攻击主要是指通过地面、空中预警指挥系统、友机之间的信息支援，进行态势通报、空域分配、目标分配等，使作战机群飞机之间能够相互协调地进行各自的多目标攻击，提高机群作战效能。第三代战斗机已具备多机协同单目标攻击能力。第四代战斗机在此基础上，将之进一步发展为多机协同多目标攻击。

美国的远期国防规划《空军2025》认为，未来的作战飞机编队将是由多种飞机组成的混合机队。编队飞机之间、作战飞机和地面之间、作战飞机和空中预警机之间都将通过全球通信网络进行相互通信，多机协同多目标攻击将在更大的战场范围中进行。

（3）空空反辐射导弹攻击

其典型代表为俄罗斯的X-31反预警机型中的空空反辐射导弹。该弹采用惯性中制导及被动/主动复合雷达末制导，装有冲压喷气/固体火箭组合式发动机，在加速到一定速度后抛去固体推进剂火箭助推器，然后借助喷气发动机进行续航。该弹的射程达到200 km，用于攻击远距离非机动飞机，如预警机等。

8.5.3.2 智能机载武器火力指挥控制技术

智能机载武器火力指挥控制（IA WFCC）系统是利用人工智能（AI）技术将机载武器火力指挥控制问题中的传感器数据融合、目标信息分析、敌我战术态势评定、武器资源调度、火力效能评估、攻击轨迹控制等技术进行智能化设计，并且以专家系统、神经网络、模糊控制等为实现基础，从而能够应对现代空战日趋严峻的不确定性和复杂性。它的引入将大大地提高机载武器火力指挥控制系统的作战效能，并且具有极强的系统自适应能力和人机交互性，从而能够推动机载武器火力指挥控制系统总体技术的发展。

1. 智能机载武器火力指挥控制系统的构成

新型的机载武器火力指挥控制系统利用人工智能技术对火力/飞行耦合控制器进行了智能化设计，成功地实现了“仿人智能火力/飞行耦合控制器”，通过空空和空地机炮两种攻击模态的数学仿真试验证明：仿人智能火力/飞行耦合控制器有效地模拟了飞行员的控制和操纵行为，充分地满足了火力/飞行控制的系统总体要求，对控制环境的变化具有极强的自适应能力和鲁棒性。智能机载武器火力指挥控制系统的构成框图如图8.22所示。

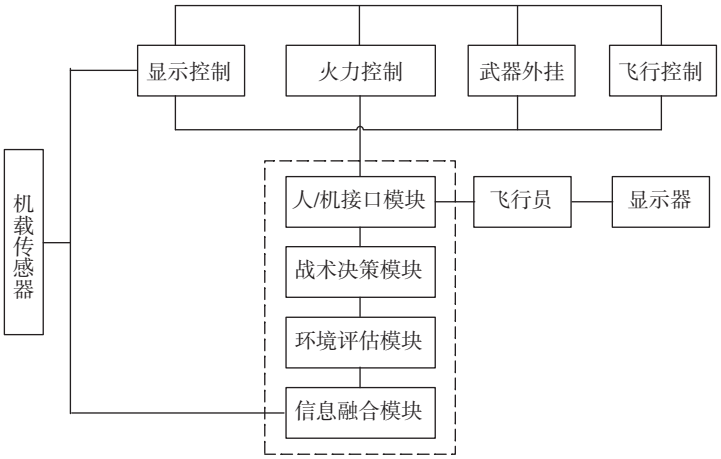


图8.22　智能机载武器火力指挥控制系统构成框图

在图8.22中，虚框内即为智能机载武器火力指挥控制系统部分。

(1) 信息融合模块是由以下5个子模块组成的。

- 时间对准。带有状态矢量和误差协方差矩阵的目标航迹数据库将被“对准”到融合修正时间。
- 坐标校准。传感器数据（包括目标航迹数据库）需要参照共用坐标原点，对传感器偏差提供补偿。
- 组合。利用特征测试、运动测试和概率测试对来自各个不同传感器数据库的航迹和报告进行比较，以确定“融合”的备用值。
- 相关。组合结果再次被处理，以确定将要融合的航迹。
- 航迹修正。“最佳相关”的航迹信息用来修正相应的状态矢量和误差协方差矩阵，其结果是与环境评估模块相交联的目标航迹数据库。

(2) 环境评估模块是由以下3个子模块组成的。

- 敌我占位态势评估。在我机搜索、跟踪范围内，进行关于敌我双方在数量、位置和运动趋向的整体态势评估。
- 目标威胁态势评估。在我机武器火力范围内，由目标集的特征、占位及预测的运动趋向来评估各单个目标对我机的威胁态势。
- 我机火力效能评估。在我机武器火力范围内，由目标数量、占位、运动趋向以及我机武器外挂的变化情况来评估我机的火力效能。

(3) 战术决策模块是由以下3个子模块组成的。

- 战术任务决策。由环境评估信息和系统状态信息来衡量我机占据优势或受到威胁的程度，进而对我机的行动意图（攻击/干扰/躲避/保持）做出决策。
- 战术攻击决策。由战术任务决策、环境评估信息和系统状态信息来决策我机的攻击方式，其中包括武器调度、武器/目标匹配、火力控制任务机相应的武器发射条件解算和显示模块控制、武器发射指令时序控制等子功能模块的激活。
- 战术飞行动作决策。由战术攻击决策、战术任务决策、环境评估信息和系统状态信息来决策我机的飞控动作，其中包括战术躲避轨迹、机动攻击轨迹、占位态势修正耦合控制、火控/飞行误差修正复合控制等子功能模块的激活。

(4) 人/机接口模块是由以下3个子模块组成的。

- 状态切换。为智能机载武器火力指挥控制系统与飞行员之间提供平滑的状态切换。
- 信息发送。为智能机载武器火力指挥控制系统和飞行员的决策信息提供通道。
- 告警。在飞行员决策状态下，当飞行员的决策与智能机载武器火力指挥控制系统的决策差异较大时，提供警告信息。

2. 智能机载武器火力指挥控制系统的典型应用

根据世纪末的几次局部战争表明，空对地攻击已经成为现代战争的主要作战手段之一，它不仅是决定战争胜负的关键性力量，而且已成为局部战争的主角。空对地攻击能力的强弱是空中力量由积极防御型向攻防兼备型转化的决定性因素，世界各国都以加强空对地攻击力量作为今后发展空中力量的重点。因此，美国又推出了配有先进航空电子系统（“宝石台”航空电子系统）的联合攻击机F-35（见图8.23）。

图8.23 F-35战斗机

众多高新技术在F-35上汇聚，将使F-35挂上“世界最先进”的光环。F-35先进综合航空电子结构的主要特点，是在综合式航空电子结构的基础上，采用了整个系统统一的航空电子网络，并进一步推进了传感器系统综合。

F-35技术特点如下。其火控系统不仅具有功能强大的综合核心处理机，而且还具有高度综合的传感器部分。与此同时，还配备了更为先进的机载AESA多功能雷达，综合高效的电子战系统，友好的人机界面——下视显示器和头盔显示器，以及综合完善的通信、导航、识别系统。高度可靠的飞机管理系统使F-35具有更强的通用性，能在全球任何基地和航母上起降，与F/A-22形成高/低档搭配，完成战术武器投放、近距离空中支援、纵深打击等对地攻击任务。

8.6 无人系统

自从有了战争，人类就一直在努力寻找战场上冲锋陷阵的替身。无人作战将部分取代有人作战功能，人机协同作战与体系对抗将在未来信息化战争中更为突出。

军用无人系统的发展恰恰适应了这种军事需求的转变，相关技术的发展又为发展军用无人系统提供了必要的物质技术基础。无人机、无人潜水器和军用机器人这样的无人化作战平台将越来越受到重视，无人化武器装备的种类将得到新的发展。未来战争中，担任突击任务的将是由“机器人战士”组成的集机械化、信息化、电子化、机动化和隐身化为一体的无人化部队，它将成为未来战争的一支重要力量，未来战争将可能出现无人化战场。

军用无人系统具有智能化程度高、作战使用灵活、综合作战效益高、适用于危险环境、人员伤亡率低、全寿命费用低等突出特点，特别适用于信息化战争的非接触、非线性、非对称远程打击等要求。无人智能作战系统将承担最危险、高频率、长时间、高强度的作战任务，但终究智能无人系统还是由人来操纵的，是人在战场上的延伸。

8.6.1　机器人

机器人也称为地面无人系统，它可以代替人在高度恶劣、危险的环境下进行复杂而精细的工作。机器人是机械技术与控制技术的完美结合，在工业制造、资源勘探以及国防安全方面发挥着巨大的作用。

8.6.1.1　机器人的分类

机器人种类繁多，可以从不同的角度对其进行分类，如机器人的发展程度、结构形式、控制方式、信息输入方式、智能程度、应用领域等。因此，目前仍是仁者见仁，智者见智，国际上并没有制定统一的标准。

我国的机器人专家从应用领域出发，将机器人分为两大类，即工业机器人和特种机器人。工业机器人就是面向工业领域的多关节机械手或多自由度机器人。而特种机器人则是除工业机器人之外的，用于非制造业并服务于人类的各种先进的机器人，包括军用机器人、水下机器人、空间机器人、娱乐机器人、微小型机器人与微操作机器人等。

1. 军用机器人

军用机器人是一种用于军事领域的具有某种仿人功能的自动机，从物资运输到搜寻勘探以及实战进攻，军用机器人的使用范围非常广泛。

2. 水下机器人

水下机器人是指在海洋深水或在江河湖泊浅水中进行观察、作业的机器人。

3. 空间机器人

空间机器人是指开发太空资源，进行空间建设和维修，协助空间生产和科学实验，以及星际探索等方面的机器人。

4. 娱乐机器人

娱乐机器人以供人观赏、娱乐为目的，具有机器人的外部特性，可以像人、像动物、像童话或科幻小说中的人物，具有语言能力，会唱歌，有一定的感知能力。

5. 微小型机器人与微操作机器人

微小型机器人以其体积小，成本低，以及能够在极端和狭小的空间内协作作业等特点，被应用在军事、民用等特殊的场合。而微操作机器人则采用机器人技术进行微细操作。

8.6.1.2　典型装备

这里介绍几种军用机器人，它们主要担负部队中最危险、最单调的任务。

1. 遥控爆炸物销毁系统（RONS）

美国海军研制的遥控爆炸物销毁系统（见图8.24）是联合机器人总计划的一个项目。该机器人是一辆履带式遥控车，装备有CCD摄像机、无线电、光纤通信设备及三轴腕关节的七功能机械手，使用柴油机及蓄电池驱动，可爬越45° 角楼梯和6 m高的45° 角斜坡。

图8.24　RONS系统

2. 无人侦察车

美军系列无人侦察车是一个大的家族，比较出名的有“帕克波特”（Pack Bot）、“城市勇士”（MATILDA）、“角斗士”（ARV-A-L）无人侦察车等。以“城市勇士”无人侦察车为例，它装备了昼/夜照相机、彩色显示器、录音设备，以及一些即插即用的侦察设备，如核生化传感器或其他光电系统，主要用于远距离侦察、检查、评估和取样，甚至可进入洞穴、地道、下水道系统进行侦察。

而“角斗士”无人侦察车装备了多光谱三维成像的日/夜摄像机、毫米波雷达、激光测距仪、生化武器探测设备和攻击性武器系统等，可以在任何天气与地形下，执行昼夜侦察、核生化武器探测、突破障碍、反狙击手和直接射击等任务。它的侦察设备为后方提供战场前沿的红外、可见光图像或激光指示（见图8.25）。

图8.25 美军无人侦察车

3. AC-ROV SP50水下机器人

AC-ROV SP50水下机器人体积小，尺寸为203 mm × 152 mm × 146 mm，质量仅为3 kg，功耗也非常小，连同水面控制台，其功耗小于500 W，最大潜水深度可达75 m，具有六个推进器、能以六自由度运动，所有的推进器都位于潜水器内，能够减小ROV的水下阻力。另外，还有4个卤素灯和1个摄像头用于水下观察（见图8.26）。

图8.26 AC-ROV SP50水下机器人

8.6.2 无人机

无人机是一种利用信息技术革命成果而发展的高性能信息化武器装备，其智能化程度高，因此被称为名符其实的飞行机器人。军用无人机主要用于侦察、监视、通信中继，甚至已经可以从空中发射武器攻击地面目标。对提高战场空间感知能力、高风险目标突防能力、通信导航支援能力、电子战能力、敌防空系统压制能力、固定和移动目标攻击能力、高过载机动能力、作战生存能力、联合作战能力与主宰战场空间能力等，起着举足轻重的作用，在未来战争中处于突出的地位，因此又有着“空中超人”的美称。

8.6.2.1 主要功能

1. 作为防空武器的靶机

无人机的最初用途是作为靶机使用，主要用于地面防空和空中格斗武器的试验与训练。

2. 侦察监视

无人侦察机依靠机载设备（包括可见光照相机、电影摄影机、标准或微光电视摄像机、红外扫描器和雷达等设备），深入阵地前沿和敌后100~200 km，甚至更远的距离，完成各种侦察和

监视任务。一般来说，一架无人机可携带一种或几种侦察设备，按预定的程序或地面指令进行工作，最后将所获得的信息随时传送回地面，同时也可将获得的所有信息记录下来，待无人机回收时一次取用。

3. 实施干扰

无人机可对敌人的信息系统实施干扰，使其通信中断，指挥失灵。无人机主要完成对雷达和通信的干扰，使敌方高炮和导弹阵地无法得到所需的情报信息。为此，一架无人机可同时装备两种或两种以上的干扰设备。在光电对抗中，无人机的作用潜力也十分引人注目。它可以利用装备的烟雾装置，瓦解敌方的光电制导武器的进攻；也可以装备闪光灯具，作为红外诱饵，引偏敌方的红外制导武器；同时还可以利用自身机动灵活和滞空时间长的特点，把携带的曳光弹准确地投放到所需的位置上。

4. 对地攻击

作为一种空中运载工具，无人机也能携带多种对地攻击武器，飞往前线或深入敌占区纵深，对地面军事目标进行打击；它可以用空对地导弹或炸弹对敌防空武器实施压制；用反坦克导弹等对坦克或坦克群进行攻击；用集束炸弹等武器对地面部队集结点等进行轰炸。特别值得一提的是反辐射攻击无人机，这是一种利用敌方雷达辐射的电磁波信号，发现、跟踪以至最后摧毁雷达的武器系统。这种无人机不仅可用于攻击敌方雷达、干扰机和其他辐射源，而且高速反辐射无人机在加装复合制导装置等设备后，还可用于攻击敌预警机和专用电子干扰飞机。

5. 通信中继

通信中继是无人机最具前景也最重要的功能之一。通过装载的抗干扰扩频通信设备、大功率固态放大器、全向甚高频和超高频无线电台中继设备等，可进行数据、信号、话音和图像通信。

8.6.2.2　典型装备

1. RQ-4A“全球鹰”高空无人战略侦察机

RQ-4A“全球鹰”无人侦察机以续航时间长和侦察监视能力强闻名世界，其最快飞行速度为644 km/h，最大飞行高度为19.81 km，航程可达25 000 km，续航时间可达38 h（见图8.27）。它可同时携带光电/红外传感器系统和合成孔径雷达，其中光电/红外传感器系统包括第三代红外传感器和一个柯达数字式电耦合器件（CCD），光电传感器工作在0.4~0.8 μm波段，红外传感器系统工作在3.6~4.0 μm的中红外波段；合成孔径雷达工作在X波段，该雷达获取的条幅式侦察照片可精确到1 m，定点侦察照片可精确到0.3 m。

图8.27　“全球鹰”无人侦察机

在一次飞行中，“全球鹰”无人侦察机的光电/红外传感器可提供74 000 km²范围内的高分辨率目标图像，合成孔径雷达探测距离为20~200 km。“全球鹰”无人侦察机主要有两种侦察方

式：一种是广域搜索模式，用来快速覆盖大范围区域；一种是点状搜索模式，用来重点仔细侦测小块可疑地区。一天之内，“全球鹰”若用广域搜索模式工作，可侦搜100 000 km²的区域，图像分辨率可达0.9 m；若用点状搜索模式工作，可仔细侦测1900个2 km × 2 km的可疑地区，图像分辨率高达0.3 m，而且不管白天黑夜，也不论云雨、沙尘暴等恶劣气象条件，其机载雷达都能够准确地识别大部分包括伪装在内的地面各种飞机、导弹、车辆及其他军事目标。

2. RQ-1“捕食者”无人机

“捕食者”无人机是一种能携带空地导弹的中、低空长航时无人侦察监视系统，其最快速度可达216 km/h，续航时间可达40 h以上，最大活动半径为3700 km，最大升限为7620 m。“捕食者”无人机装备了合成孔径雷达、电视摄像机和前视红外传感器等侦察设备，主要用于侦察监测。其中，可变光圈的电视摄像机主要用于白天侦察，可变光圈的前视红外摄像机主要用于暗光和夜视条件下的拍摄。合成孔径雷达的扫描范围为800 m × 10.8 km，图像分辨率可达0.3 m，具有全天候的侦察监视能力。因此，“捕食者”无人机在浓雾、雨雪、黑夜、烟幕等复杂气象条件下也可进行拍摄，其拍摄的各种图像可通过数据链或卫星通信系统，实时传给后方。

3. 小型/便携式无人机

FQM-151“短毛猎犬”无人机、RQ-11A“渡鸦”无人机、“美洲狮”无人机、“龙眼”无人机和“黄蜂”微型无人机等型号，都是小型便携式无人机，轻则几kg，重则超过20 kg，可以装在特种背囊中随身携带，甚至可手持发射起飞，留空时间从几十分钟到几小时不等；其传感器多为电视摄像机和红外摄像机，主要用于小分队对战场前沿的监视侦察、目标指示、火力校射和损失评估等。而MQ-9“死神”无人机、MQ-5B“猎人”多用途无人机、RQ-170“哨兵”隐形无人侦察机和RQ-8“火力侦察兵”无人侦察直升机，在美国陆军的侦察体系中同样也占有甚至更重要的地位。其中，MQ-9“死神”无人机装备了电子光学设备、红外系统、微光电视、合成孔径雷达以及“地狱火”空地导弹和500磅激光制导炸弹，具有很强的侦察监视能力和对地攻击能力；MQ-5B“猎人”多用途无人机可在4572 m的空中连续飞行21 h，其主要传感器有电视摄像机和前视红外摄像机；RQ-8“火力侦察兵”无人侦察直升机则携带了合成孔径雷达、光电及激光传感器和激光指示器及测距仪，可以为作战部队提供极其精确的目标信息。

思考题

1. 指挥自动化系统是什么？可以分为哪几部分？
2. 情报预警的方式共分为哪几种，它们有什么区别？
3. 天基与空基侦察方式有什么不同？
4. 红外探测设备是什么？
5. 什么是数据链？
6. 试介绍火控计算机的主要功能及组成。
7. 超视距多目标攻击的火控关键技术包括哪几种？
8. F-35火控系统的技术特点是什么？
9. 无人系统是什么？无人机的主要功能是什么？

参 考 文 献

[1] 张训才、苏宣. 军事信息技术. 北京：解放军出版社，2007.
[2] 解放军理工大学编著. 军事信息技术概论. 北京：军事科学出版社，2010.
[3] 张官海、魏长智. 军事信息技术基础. 北京：蓝天出版社，2006.
[4] 童敏明、戴新联. 现代传感器技术. 徐州：中国矿业大学出版社，2006.
[5] 赵琳、丁继成、马雪飞. 卫星导航原理及应用. 西安：西北工业大学出版社，2011.
[6] 籍宝林、杨飞龙、闫科. 数字化部队及作战运用研究. 北京：解放军出版社，2011.
[7] 吕翊. 电信传输技术. 北京：清华大学出版社，2011.
[8] 罗国明，沈庆国，张曙光. 现代交换原理与技术. 北京：电子工业出版社，2010.
[9] 周荣庭. 军事信息管理概论. 军事科学出版社，2001.
[10] 肖占中、安虎成、刘承红. 军事信息管理. 解放军出版社，2005.
[11] 王淑江，刘晓辉. 网络存储·数据备份与还原. 电子工业出版社，2010.
[12] 边肇祺、张学工. 模式识别. 清华大学出版社，2000.
[13] 洪文学等. 基于多元统计图表示原理的信息融合和模式识别技术. 国防工业出版社，2008.
[14] 杨静宇、邬永革、刘雷健等. 战场数据融合技术. 兵器工业出版社，1994.
[15] 康耀红. 数据融合理论与应用. 西安电子科技大学出版社，2006.
[16] David L. Hall，James Llinas. 多传感器数据融合手册. 杨露菁、耿伯英，译. 电子工业出版社，2008.
[17] 李弼程、黄洁、高世海、王天鹏. 信息融合技术及其应用. 国防工业出版社，2010.
[18] 陈振宇、曹婉. 战场环境与可视化技术. 军事科学出版社，2006.
[19] 张涛、曹婉、陈振宇. 战场环境与可视化技术. 军事科学出版社，2008.
[20] Robert Spence. 信息可视化交互设计. 陈雅茜，译. 机械工业出版社，2012.
[21] 刘熹、张寒、刘海燕、田畅. 战场态势一致性技术专题讲座（一）. 军事通信技术，2012.12
[22] 陈燕. 数据挖掘技术与应用. 清华大学出版社，2011.
[23] 赵亮、萧德云、刘震涛. 一种用于挖掘正负关联规则的可量化标准. 计算机工程，2007.2
[24] 于代军、李莉、傅亮. 军事信息资源分类组织研究. 国防大学出版社，2012.
[25] 军事信息资源分类法编委会办公室. 军事信息资源分类法使用手册. 军事科学出版社，2008.
[26] 军事信息资源分类法编委会办公室. 军事信息资源分类法. 军事科学出版社，2005.
[27] 陈围、王自华. 军事信息资源检索. 国防大学出版社，2008.
[28] 李洁、王兆勇. 网络军事信息资源检索概论. 军事谊文出版社，2011.
[29] 杨向明. 网络时代信息推荐技术及相关问题. 江西图书馆学刊，2005.
[30] 彭国莉. 信息推送技术与信息推送服务. 信息技术，2001.8

[31] 鄢朝晖、方宜仙. 个性化信息服务的新形式——论信息推拉服务. 吉林大学学报，2007.5
[32] 葛嘉佳. 网络个性化信息服务综述. 计算机时代，2004.5
[33] 周明全. 网络信息安全技术. 西安：西安电子科技大学出版社，2003.11
[34] 张红族. 信息安全技术. 北京：高等教育出版社，2008.6
[35] 徐国爱. 网络安全（第2版）. 北京：北京邮电大学出版社有限公司，2007.9
[36] Douglas R. Stinson. 密码学原理与实践（第三版）. 冯登国，译. 北京：电子工业出版社，2016.1
[37] 杨波. 现代密码学（第2版）. 北京：清华大学出版社，2007.4
[38] 何大可、彭代渊、唐小虎. 现代密码学. 北京：人民邮电出版社，2009.9
[39] 张载德. 战争神经中枢：美军全球四大指挥中心国际展望. 2001.10
[40] 张红、李峰. 第四代战斗机火控系统总体技术分析. 电光与控制，2000.8
[41] 赵伟、庞思伟. 智能化自动控制系统问题. 四川兵工学报，2010.2
[42] 任邵东、郝维平、周生炳. 美军指挥自动化系统一体化建设发展研究浅议. 航天控制，2004.4
[43] 董思鹏、刘兴. 综合电子信息系统（第二版），国防工业出版社，2008.
[44] 黄烈炎、魏蛟龙. 美军数据链建设及启示. 舰船电子工程，2005.2
[45] 李雪超、张金成. C4ISR体系结构框架的新发展. 现代防御技术，2011.2
[46] 周彬. 光学侦察卫星及反侦察技术综述. 光电技术应用，2004.5
[47] 刘进军. 导弹预警卫星：刺穿时空的鹰眼. 卫星与网络，2012.10
[48] 王群. 美国新一代导弹预警卫星系统及其能力分析. 国防科技，2012.2
[49] 中国人民解放军总参谋部军训部. 军事通信基础，国防工业出版社，2012.